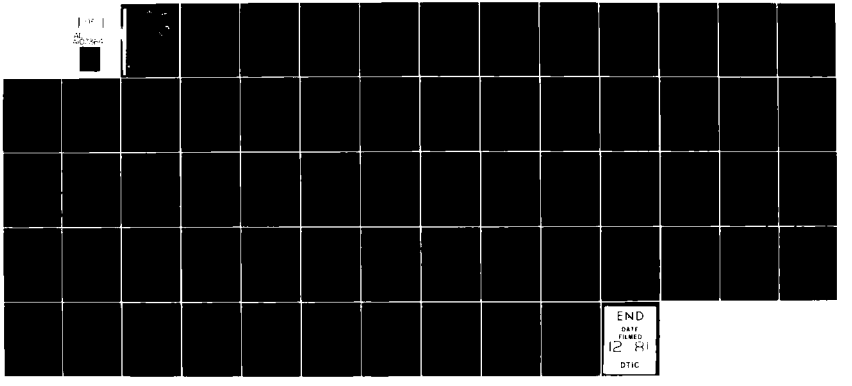
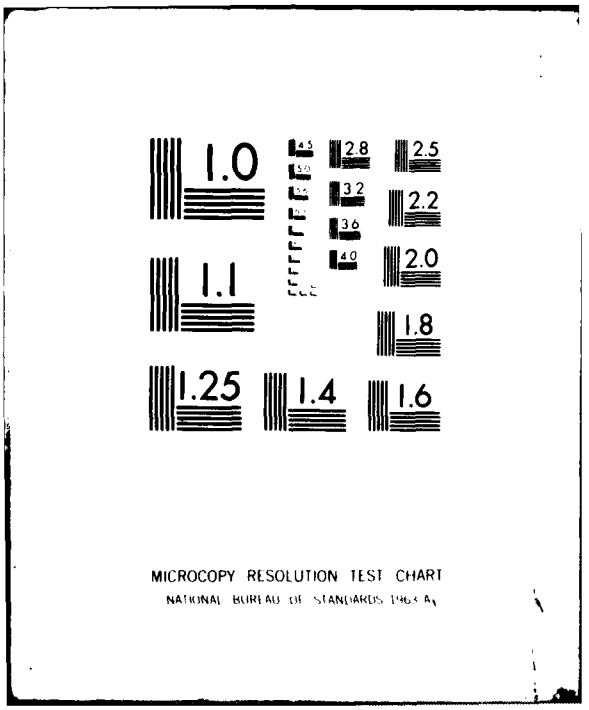
AD-A107 364 MISSOURI UNIV-COLUMBIA TAILORED TESTING RESEARCH LAB F/G 12/1
A COMPARISON OF PROCEDURES FOR CONSTRUCTING LARGE ITEM POOLS.(U)
AUG 81 R L MCKINLEY, M D RECKASE N00014-77-C-0097
UNCLASSIFIED RR-81-3 NL



END
DATE FILMED
12 81
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RR-81-3	2. GOVT ACCESSION NO.: AD-A107364
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): A Comparison of Procedures For Constructing Large Item Pools.	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Robert L. McKinley and Mark D. Reckase	8. CONTRACT OR GRANT NUMBER(s): N00014-77-C-0097
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Educational Psychology, University of Missouri, Columbia, MO 65211	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: P.E.:61153N Proj.: RR042-04 T.A.:042-04-01 W.U.:NR150-395
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Programs, Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: August 81
	13. NUMBER OF PAGES: 50
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linking Procedures
Latent Trait Models
One-Parameter Model
Three-Parameter Model

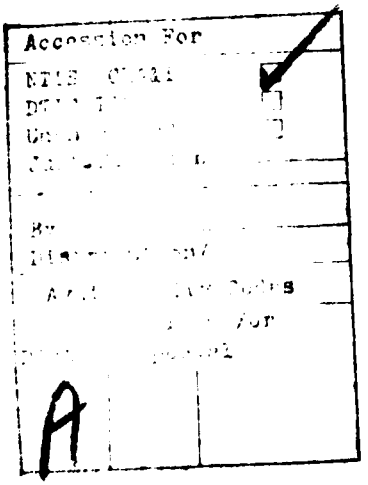
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The application of testing methodology is often aided by the development of large item pools. Also, newer testing methodologies, such as tailored testing, almost require such pools. Furthermore, the recent truth-in-testing legislation in New York State has increased interest in the formation of large item pools with item calibration information on the same metric so that high quality tests can be produced every year, despite the requirement that each test be made public after its use.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

This study was conducted in order to evaluate available linking techniques for forming large item pools and to make recommendations as to which technique should be used under various circumstances. Variables of interest included calibration model and procedure, sample size, overlap level, and linking procedure. The calibration models considered were the one- and three- parameter logistic models. The calibration procedures that were considered included the MAX calibration program for the one-parameter logistic (1PL) model, and the LOGIST and ANCILLES calibration programs for the three-parameter logistic (3PL) model. Sample sizes of 100, 300, 500, 1000, and 2000 were used with overlap levels of 5, 15, and 25 items. The linking procedures investigated included major axis linking, least squares linking, least squares with outlier deletion, and maximum likelihood linking. The basic design of the study was to sample short tests from a longer test in such a way that each short test selected had a predetermined number of items in common with the test just previously sampled. For each short test response data were obtained for a subset of the large number of examinees for which response data from the longer test were available. The short tests were then calibrated and linked using the linking methods selected for the study. The resulting parameter estimates were then compared to the estimates obtained from a calibration of the full sample for the longer test, which served as the criteria by which the linking procedures were evaluated. Response data used for this study were for a sample of 4000 examinees from an administration of the Iowa Tests of Educational Development during the 1975-1976 school year. From the results of the analyses performed on these data the following conclusions were reached. For the best results an overlap of 15 items appeared to be best. At the 15 item overlap level a sample size of 2000 appeared to be necessary for stable linking of the 3PL model parameters, although when LOGIST was used 1000 seemed to be a sufficient sample size for linking item discrimination estimates. For the 3PL model the LOGIST program appeared to yield the best overall results. With a sample size of 2000 all of the linking procedures yielded adequate results. For the 1PL procedure a sample size of 100 to 300 appeared to yield adequate results.



CONTENTS

A Comparison of Procedures For Constructing Large Item Pools

The application of testing methodology is often aided by the development of large item pools. For example, traditional test construction is facilitated by the presence of large item pools from which items can be selected. Moreover, if the item analysis information available for all of the items in the pool is on the same metric, comparisons can be made between the technical quality of the items. Although traditional test construction is facilitated by the availability of such pools, the newer testing methodologies of computer assisted test construction and tailored testing almost require such pools. The recent truth-in-testing legislation in New York State has also increased interest in the formation of large item pools with item calibration information on the same metric so that high quality tests can be produced every year, despite the requirement that each test be made public after its use. It is the purpose of this report to review and evaluate the techniques for forming large item pools and to make recommendations as to which technique should be used under various circumstances.

Logically, one can conceive of several procedures for calibrating large item pools. The simplest of these is to produce a test the size of the desired item pool and administer this test to a large sample of individuals. The large sample is required so that stable estimates of the many item parameters can be obtained. Although this procedure can be used in some situations, it is often impossible to get the large sample size and long testing time needed to administer such tests. Also, if it is desirable to increase the size of the item pool at some later date, this procedure would require that an even longer test be produced and administered.

A modification of this procedure which solves some of these problems is to make up a number of distinct, shorter tests and administer these tests to the same group over a number of testing sessions. This procedure removes the requirement of a single long testing session, but adds problems concerning examinee attrition and possible changes in the ability of the sample over the series of testing sessions.

A second alternative to the single long testing session is to administer several short tests to a number of comparable samples. These samples could be obtained by randomly sampling from the same population, or by matching examinees in each sample. This procedure also removes the requirement for a single long testing session, but in turn adds problems in matching or in obtaining equivalent random samples from the same population. Thus, although each of these procedures could be used to produce a large item pool, each has disadvantages that make it impractical for most situations.

Another alternative to the three procedures listed above has the capability of solving most of the practical problems. This procedure involves administering a number of short tests to separate groups of individuals, with the requirement that each of the tests have items in common with at least one of the other tests administered. The item parameter estimates obtained on these common items are used to determine transformations that can be used to place all parameter estimates on the same scale. This procedure has the advantages of using short tests and easily obtained samples

-- no matching or random sampling is required. However, it is crucial with this procedure that accurate item parameter transformations be obtained. An evaluation of the procedures for obtaining these transformations is the main topic of this report.

The methodology for obtaining item pools using this last procedure has been given a special label because of its common usage. The technique for putting the item parameters on the same scale is called linking because the sets of common items between tests are called calibration links. For this paper linking is defined as a technique, based upon items that are in common between tests, used to put item parameter estimates obtained from different samples on the same scale of measurement.

A distinction must be made here between item parameter linking and vertical equating. Although the transformations obtained for vertical equating are somewhat similar to those obtained for linking, the purpose of the two procedures are different. Whereas linking attempts to place all of the item parameter estimates on the same scale, vertical equating attempts to put ability estimates from tests of different difficulty on the same scale. The result of this different orientation is a difference in the desired precision of the various estimates. Linking requires precise item parameter estimates, while vertical equating requires precise ability parameter estimates. It is interesting to note that vertical equating has long been a concern of the educational community, but that linking has only recently become of interest.

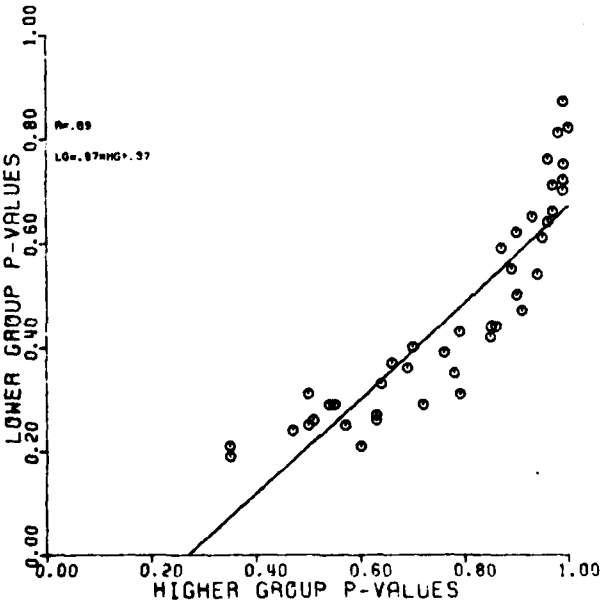
The specific problem to be addressed by this report is how to best link item calibrations to form large item pools. Although linking is conceptually simple, requiring only the development of a transformation based on the common items to get the item parameter estimates on the same scale, numerous variables affect the quality of the linking. Furthermore, the optimal combination of values for these variables has not been determined. The variables identified as being of interest for this report include: (a) the item analysis model, (b) the linking procedure, (c) the item calibration program, (d) the sample size required for stable linking, and (e) the number of common items required. These variables were manipulated in the research reported here to determine their effect on linking accuracy and on the drift of item parameter estimates from their expected values when tests are repeatedly linked. Before reporting the present study a discussion of the available item analysis models and linking procedures will be presented.

Item Calibration Models

In theory, parameter estimates obtained using any item calibration procedure can be linked to form item pools with parameter estimates on the same scale. This holds for traditional item parameters, such as proportion correct (item difficulty) and item-test correlation (item discrimination), as well as for the more recently developed item parameters from latent trait theory. As long as the same ordinal arrangement of item parameter estimate magnitudes is maintained across administrations of a set of common items, the parameter estimates can be put on the same scale.

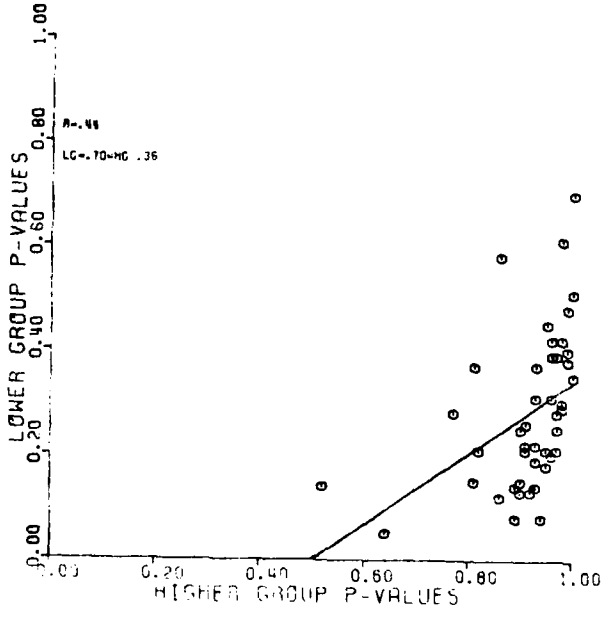
Suppose, for example, we wished to link the difficulty parameter estimates obtained from a traditional item analysis. Two 100 item tests could be produced so that fifty items were in common, and the two tests could then be administered to different groups. In order to emphasize that the characteristics of the groups are unimportant, suppose that the mean scores for the two groups differ by two standard deviations. In order to perform the linking, a transformation must be determined that will convert the one set of parameters onto the same scale as the others. Figure 1 shows a plot of the proportion of correct responses to the 50 common items obtained from two simulated groups of examinees. The range of difficulty values for the items for the higher ability groups ranged from .35 to 1.00, while the range for the lower ability group for the same items was .19 to .87. Note that despite the difference in difficulty of the 50 items for the two groups the plot still fell relatively close to a straight line and the correlation between the parameter estimates was .89. A regression equation was easily determined to predict the low ability group's difficulty values (LG) from those of the high ability group (HG), yielding LG=.87 x HG + .37. Thus the low ability group's difficulty parameters for all of the 100 items in their test could easily be converted to the high ability group's scale using this equation. However, from Figure 1 it is apparent that some curvilinearity exists in the relationship of the two sets of estimates. This curvilinearity can become a serious problem .

FIGURE 1
PLOT OF ITEM DIFFICULTIES OBTAINED
FOR TWO GROUPS TWO STANDARD
DEVIATIONS APART IN ABILITY



Although the linking of the traditional proportion correct could be performed fairly easily for these two tests, the linking may be more difficult in other situations. For example, if the difference in the ability of the two groups used to calibrate the tests is more extreme, say four standard deviations apart, the relationship between the two sets of parameter estimates is clearly curvilinear and finding an appropriate transformation is more difficult. A plot of the proportions correct for the common items on two 100-item tests for this situation is presented in Figure 2 to demonstrate this effect. The regression equation for these data is LG=.70 x HG - .36, but its inaccuracy is clearly seen from the straight regression line on the curvilinear scatter plot.

FIGURE 2
PLOT OF ITEM DIFFICULTIES OBTAINED
FOR TWO GROUPS FOUR STANDARD
DEVIATIONS APART IN ABILITY



It is not surprising that the relationship between the proportions correct for the items would be curvilinear in this case. The p-values are restricted to the range of 0 to 1, while the difference in ability of two groups can be quite extreme. The relationship must "bend" to fit into the finite range alloted to both sets of parameters.

Plots of the discrimination values -- point biserial correlations -- corresponding to the data shown in Figures 1 and 2 are presented in Figures 3 and 4. From these figures it can be seen that there is a low negative correlation between the point biserial correlations, and in Figure 3 the relationship appears to be curvilinear. The curvilinearity of the relationships and the low negative correlations would make linking of point biserial correlations quite difficult.

FIGURE 3
PLOT OF ITEM DISCRIMINATIONS OBTAINED
FOR TWO GROUPS TWO STANDARD
DEVIATIONS APART IN ABILITY

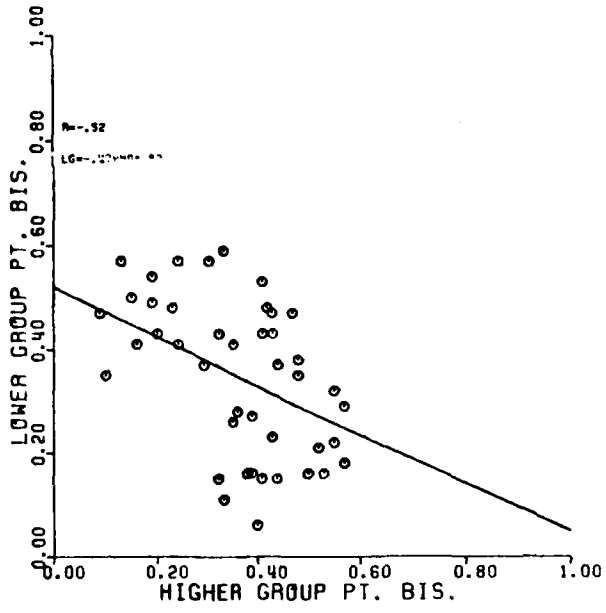
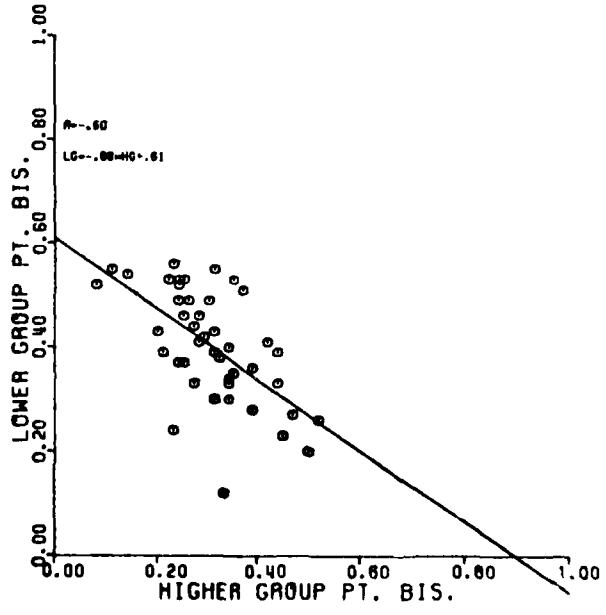
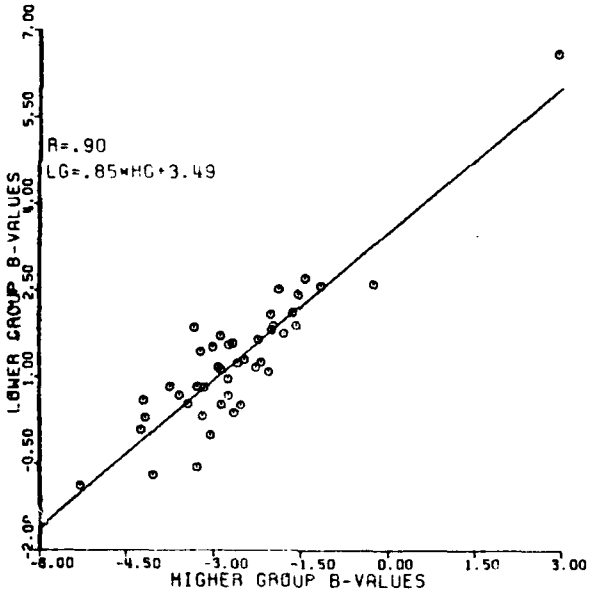


FIGURE 4
PLOT OF ITEM DISCRIMINATIONS OBTAINED
FOR TWO GROUPS FOUR STANDARD
DEVIATIONS APART IN ABILITY



In recent years, alternatives to the traditional item analysis techniques have been developed that are based upon probabilistic models of the interaction of a person with an item. These models, called latent trait or item response theory (IRT) models, do not restrict the possible range of the item parameters and therefore will often yield linear relationships between sets of parameter estimates when traditional item statistics do not. To demonstrate this fact, the plots of the IRT difficulty parameter and discrimination parameter estimates for the two groups differing by four standard deviations are shown in Figures 5 and 6, respectively. Note that the curvilinearity present in Figures 2 and 3 is not present when this alternative model is used. Because of the convenience brought about by this linearity, the evaluation of linking techniques reported here will concentrate on those techniques used with two of the more commonly applied IRT models. These two models are described below.

FIGURE 5
PLOT OF ITEM DIFFICULTIES OBTAINED
FOR TWO GROUPS FOUR STANDARD
DEVIATIONS APART IN ABILITY



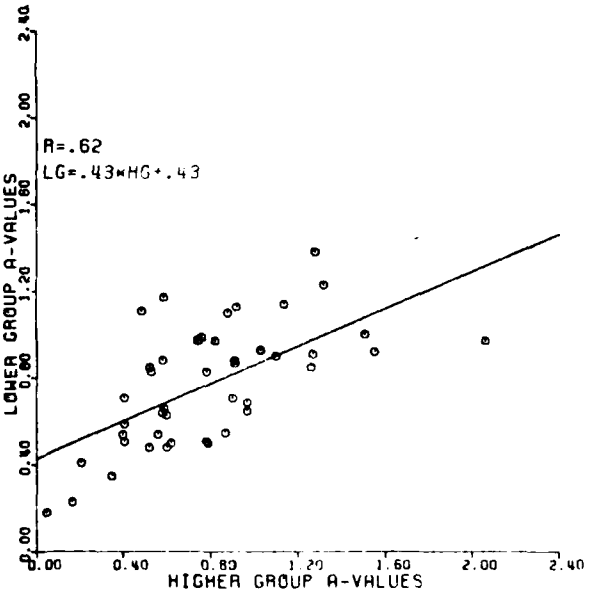
One-Parameter logistic model

The one-parameter logistic (1PL) IRT model was intially developed by Rasch (1960). The model is given by the equation

$$P(u_{ij}) = \frac{e^{u_{ij}(b_i - \theta_j)}}{1 + e^{(b_i - \theta_j)}} ,$$

where u_{ij} is the score obtained by Person j on Item i, $P(u_{ij})$ is the probability of the item score, b_i is the difficulty parameter for Item i, and

FIGURE 6
PLOT OF ITEM DISCRIMINATIONS OBTAINED
FOR TWO GROUPS FOUR STANDARD
DEVIATIONS APART IN ABILITY



θ_j is the ability parameter for Person j. This model assumes a unidimensional latent trait or, equivalently, local independence. It also assumes no guessing and equal discrimination for all items. Although these assumptions are not reasonable for many tests (i.e., multiple-choice achievement tests), the model has been used with some success for a number of applications (Rentz and Bashaw, 1977; Ireland, 1976; Woodcock, 1972).

Since the estimation of the parameters of the IRT models requires lengthy and sophisticated statistical procedures, parameter estimation is usually performed through the use of a computer program. The quality of the parameter estimates is therefore dependent on the program used for estimation. Thus, in order to fully describe the procedure, the program used for parameter estimation must be discussed in addition to the model. For the analyses described in this report the parameter estimates for the 1PL model were obtained using a modified version of the MAX program developed by Wright and Panchapakesan (1969).

Three-parameter logistic model

The three-parameter logistic (3PL) IRT model was developed by Birnbaum (1968) as a more mathematically tractable substitute for the normal ogive model (Lord, 1952). The model is given by the equation

$$P(u_{ij} = 1) = c_i + (1 - c_i) \frac{e^{Da_i(b_i - \theta_j)}}{1 + e^{Da_i(b_i - \theta_j)}} ,$$

where u_{ij}, $P(u_{ij})$, b_i, and θ_j are as defined above, D is the constant 1.7, used to increase the similarity of the logistic function to the normal ogive function, c_i is the lower asymptote of the item characteristic curve for the item (pseudo-guessing level), and a_i is the discrimination parameter of the item. This model also assumes a unidimensional latent trait or, equivalently, local independence. This model has been used for numerous testing applications by the Educational Testing Services (Marco, 1977).

For the analyses described in this report, two different calibration programs were used to obtain parameter estimates for the 3PL model. These were the LOGIST program mentioned above, and the ANCILLES program developed by Urry (Croll and Urry, 1978). The ANCILLES program differs from LOGIST in that it uses a combination of minimum chi-square and ancillary estimation procedures instead of maximum likelihood. Since these techniques yield somewhat different parameter estimates, it was felt to be important that both be used to evaluate the linking procedures.

Linking Procedures

Several procedures for linking calibrations have been identified in the literature, and several other procedures have been developed on this project. They include: (a) the major axis method, (b) the least squares method, (c) the least squares method with outlier deletion, (d) the maximum likelihood method, and (e) the ICC equating method. The first four of these procedures are based on the item difficulty parameter scale, while the last procedure is based on the ability scale. The first four procedures were used in this study, while the fifth was not. The four procedures used in this study will be described first. Then the last procedure, the ICC equating method, will be discussed as a possible alternative to the procedures currently in use. It was not used in this study because it was not available at the time this study was conducted.

Major Axis Method

The major axis method transforms the parameter scales for the items from the various tests using equations that correspond to the equations for the major axes of the ellipses of the scatter plots formed by the parameter estimates from two administrations of a set of items. When the 1PL model difficulty parameter estimates are being linked, the following procedure is followed. First, the mean difficulty parameter estimate, $\bar{b}_B$, for the linking base test, the test used to define the parameter scale, is computed for the items in common between the base test and the test to be linked. Next, the mean difficulty parameter estimate for the tests to be linked, $\bar{b}_l$, is computed for the common items. The linking constant used to determine the linking transformation is then computed by

$$k = \bar{b}_B - \bar{b}_L. \tag{1}$$

This constant is then added to all of the difficulty parameter estimates for the items on the test to be linked to give transformed parameter estimates. That is,

$$b'_{Li} = b_{Li} + k, \tag{2}$$

where b_{Li} is the untransformed difficulty parameter estimate for Item i of the test to be linked and b'_{Li} is the transformed estimate. The transformed values are then combined with the base test values for the common items using a weighted averaging procedure given by

$$b'_{Bi} = \frac{N_B b_{Bi} + N_L b'_{Li}}{N_B + N_L}, \tag{3}$$

where N_B and N_L are the sample sizes of the base test and test to be linked, respectively. This last step is performed in an attempt to increase the precision of the parameter estimates as more test results are obtained. Note that the parameter estimates for the common items are based on the combined samples, while the remaining items are based on the samples for their respective tests.

The major axis linking procedure for the 3PL model (Marco, 1977) is somewhat more complicated, partly because of the added parameters and partly because fewer assumptions are made regarding the relationship between the parameter estimates from different tests. When the 1PL model was used the assumption was made that, although the points selected as the origin might be different for different tests, the parameter estimates for different tests had the same unit of measurement. That assumption is not made when the 3PL model is used. Therefore, linking with the 3PL model must include an equating of the unit, as well as the origin. The first step, then, is to compute both the mean and the standard deviation of the difficulty parameter estimates of the common items for the base test, $\bar{b}_B$ and s_{bB}, and the test to be linked, $\bar{b}_L$ and s_{bL}. To transform the difficulty parameter estimates of the test to be linked to the scale of the base test, the following equation is used:

$$b'_{Li} = \frac{s_{bB}}{s_{bL}} \cdot b_{Li} - \bar{b}_L \cdot \frac{s_{bB}}{s_{bL}} + \bar{b}_B. \tag{4}$$

To transform the discrimination parameter estimates, the following equation is used:

$$a'_{Li} = \frac{s_{bL}}{s_{bB}} \cdot a_{Li}. \tag{5}$$

The guessing parameter estimates do not have to be transformed since the two sets of guessing parameter estimates are already on the same scale. The transformed difficulty and discrimination values for the common items

are then combined with the base test values using the same weighted averaging procedure used for the 1PL model.

Least Squares Method

An assumption that is implicit in the major axis method is that two sets of estimates for the same items have a correlation of 1.0. That is, the major axis method assumes that the true parameters of the common items are obtained from the calibration of both tests, and that the parameter estimates from the two calibrations are different only in scale. The least squares method, on the other hand, assumes that the parameter estimates contain error that needs to be controlled. That is, the assumption is made that the correlation between the two sets of estimates for the common items is less than 1.0, and that a regression procedure is needed to minimize the error introduced into the estimates of the remaining items during transformation. The first step in the least squares linking method is the computation of the slope and intercept of the regression equation predicting the base test difficulty parameter values for the common items from the difficulty values of the common items from the test to be linked. The difficulty parameter estimates of the test to be linked are then transformed to the base test scale using the regression equation. That is,

$$b'_{Li} = \beta_0 + \beta_1 b_{Li}. \tag{6}$$

Estimates for the common items are then combined using the weighted averaging procedure previously discussed.

To transform the discrimination parameter estimates the means of the a-values for the common items, $\bar{a}_B$ and $\bar{a}_L$, are computed. The a-values of the test to be linked are then transformed by the equation

$$a'_{Li} = \frac{\bar{a}_B}{\bar{a}_L} \cdot a_{ij}. \tag{7}$$

As with the difficulty values, the a-values for the common items are combined using a weighted averaging procedure. As was the case with the major axis method, when using the least squares method the guessing parameter estimates are not transformed.

Least Squares Outlier Deletion Method

The outlier deletion version of the least squares method was developed because in small sample calibrations the b-values obtained for the 3PL model are occasionally well outside the expected range (e.g., b=-32.076). It was hoped that by deleting these poorly estimated parameters from consideration in computing the linking equations the overall quality of the linking would be improved.

The actual linking procedure for the least squares outlier deletion method is the same as the least squares method with the exception that any common items with b-values more than two standard error of estimates away from the regression line are considered to be not in common. A new regres-

sion line is computed using the remaining common items. The outlier items are not deleted from the item pool, but are transformed in the same manner as the noncommon items.

Maximum Likelihood Method

The maximum likelihood linking method requires the use of the LOGIST program developed by Wood, Wingersky, and Lord (1976). In order to perform linking using this program, the test data for the tests to be linked must first be edited into a single matrix, with the columns representing items and the rows representing examinees. An example involving two seven item tests with 10 examinees each is shown in Figure 7. In this example the first 10 examinees took items one, three, four, five, six, nine, and ten. Examinees 11 through 20 took items one, two, four, six, seven, eight, and nine. Thus, items one, four, six and nine were in common to the two tests. Items not included in the test taken by an examinee are coded as "not reached" for that examinee. The matrix may be extended to include any number of tests and examinees, limited only by available computer storage. Once all of the data from the tests to be linked are edited into this matrix, the LOGIST program is run using the matrix as input, yielding maximum likelihood estimates for the parameters on all of the test items. Since there are items in common for all the pairs of tests, the parameter estimates obtained from the LOGIST program are all on the same scale.

It is important when performing maximum likelihood linking that items not included in the test taken by an examinee be coded as "not reached" for that examinee, rather than as "omitted". The LOGIST program treats "not reached" items different than actively omitted items. "Not reached" items are not included in the analysis of the data at all, while omitted items are assumed to be items for which examinees do not know the answer but could guess at the chance level. Thus, omitted responses are used in the estimation of the parameters.

ICC Equating Method

As was previously stated, this linking procedure is based on the ability scale rather than the item difficulty scale. The goal of this procedure is to equate the ICC's for the two sets of estimates to be linked by adjusting the ability scale. Using the overlapping items, two ability estimates are computed for each examinee, one for each set of estimates. Then the regression equation is computed for predicting one ability estimate, $\hat{\theta}_1$, from the other ability estimate, $\hat{\theta}_2$. This regression equation, given by

$$\hat{\theta}_1 = b_0 + b_1\hat{\theta}_2 \quad ,$$

is used to adjust the b-values, much the way the regression equation for the b-values (Equation 6) was used previously. The a-values are adjusted using Equation 7.

Figure 7

Data Matrix for Maximum Likelihood Linking With LOGIST

	1*	2	3	4*	5	6*	7	8	9*	10
1	0	NR	1	0	1	1	NR	NR	0	0
2	1	NR	1	1	1	1	NR	NR	0	1
3	0	NR	0	1	1	1	NR	NR	0	1
4	1	NR	1	1	1	1	NR	NR	1	1
5	0	NR	1	1	1	0	NR	NR	1	1
6	0	NR	1	1	0	1	NR	NR	1	0
7	0	NR	1	0	0	0	NR	NR	1	1
8	0	NR	0	1	1	1	NR	NR	1	1
9	1	NR	1	1	1	1	NR	NR	0	1
10	1	NR	1	1	1	1	NR	NR	1	1
11	1	1	NR	1	NR	1	1	1	1	NR
12	1	1	NR	1	NR	1	0	0	1	NR
13	1	0	NR	0	NR	0	1	0	0	NR
14	0	0	NR	1	NR	0	0	1	0	NR
15	0	0	NR	1	NR	1	1	1	1	NR
16	1	1	NR	1	NR	1	1	0	0	NR
17	0	1	NR	1	NR	1	1	1	1	NR
18	1	1	NR	1	NR	0	1	1	1	NR
19	1	1	NR	1	NR	1	1	1	1	NR
20	0	0	NR	0	NR	0	0	1	1	NR

* Items in common to the two tests.

Method

Two approaches may be taken to evaluate linking procedures. One approach involves the use of simulation data, while the other approach utilizes actual response data. The use of simulation data has the advantage of allowing the parameter estimates to be compared to the known values of the parameters used to generate the data. However, simulated data usually represent an unrealistic simplification of actual test results, especially in terms of factor structure and examinee variables such as guessing.

Evaluation of linking procedures using actual data has the advantage of being realistic, but without the knowledge of the true values of the parameters a good criterion for judging the adequacy of linking procedures is lacking. Because true parameter values are not known, the evaluation of linking procedures using real data involves the comparison of linked estimates with the estimates obtained from a large sample calibration or, alternatively, gauging the consistency of results from several analyses.

It is clear from this discussion that neither the use of actual data nor the use of simulated data yields a completely satisfactory evaluation technique. Because of this, the approach for the current study was made on the basis of practical considerations. The time, effort, and resource required to employ both approaches were prohibitive. Because directly applicable results were desired, the decision was made to employ actual test data to evaluate the linking procedures selected for this study.

The basic design of the current study was to sample short tests from a longer test in such a way that each short test selected had a predetermined number of items in common with the test just previously sampled. For each short test response data were obtained for a subset of the large number of examinees for which response data from the longer test were available. The short tests were then calibrated and linked using each of the linking methods selected for the study. The resulting parameter estimates were then compared to the estimates obtained from a calibration of the full sample for the longer test. Thus, the estimates obtained from the large-scale calibration served as the criteria by which the linking procedures were evaluated.

Calibration Programs

Three calibration programs were used in conjunction with the linking procedures used in this study. These included a modification of the one-parameter logistic program (MAX) developed by Wright and Panchapakesan (1967), the three-parameter logistic program (LOGIST) developed by Wood, Wingersky, and Lord (1976), and the three-parameter logistic program (ANCILLES) described by Croll and Urry (1978). The MAX program was selected since good results had been obtained with this program in the past (Reckase, 1977). The LOGIST program was selected since it had been used with success in the past, even though large samples are required for stable estimation. The ANCILLES program was used in the hope that it would yield good results with sample sizes smaller than those required by the LOGIST program.

Data

Criterion In order for the design of this study to be successfully implemented, a large sample calibration of a long test was required. Also, since the application of linking procedures to achievement testing was of interest, a test that was representative of the factor structure of a typical achievement test was desired. In order to meet these requirements,

response data for a large sample were obtained for the Iowa Tests of Educational Development, or ITED (1972). These data were obtained through Dr. William Coffman, the director of the Iowa Testing Programs.

The ITED is a general achievement measure covering seven subareas: expression, quantitative thinking, social studies, natural sciences, literature, vocabulary, and sources of information. Although this test has seven distinct subtests, a composite score is computed which correlates highly with other general measures of achievement. Also, it has a very high internal consistency reliability. A principal components factor analysis of the test confirmed its multi-dimensionality, but also indicated that the test had a strong first factor.

The ITED has 357 items, a length that made it ideal for this study. Response data were available for 4,000 examinees, including 1,000 examinees each from grades 9, 10, 11, and 12. These data were from an administration of the test during the 1975-1976 school year. The distribution of total scores on the test was negatively skewed, with a mean of 184.61 and a standard deviation of 61.51.

Tests In addition to the linking procedures and calibration programs, variables of interest included sample size requirements and the number of items in common to two tests. In order to evaluate the effects of these variables datasets of various sample sizes and varying numbers of items in common were produced. The data were then calibrated and analyzed using the calibration and linking procedures set out above.

The procedure used for developing the tests was as follows:

1. A 50 item test, designated Test A, was selected from the 355 items available (two items were discarded by the LOGIST program due to nonconvergence of parameter estimates) using a stratified random sampling scheme. The strata used were the subtests of the ITED.

2. From Test A, n items were randomly selected as common items, where n was 5, 15, or 25. Then (50-n) new items were selected from the ITED using the stratified random sampling scheme. Thus, a new test of 50 items, designated Test B, was developed so as to have n and only n items in common with Test A

3. A new test, Test C, was created so as to have n items in common with Test B using the procedure set out in Step 2. The overlap of items between Tests A and C was ignored.

4. A fourth test, Test D, was created by the same procedure so as to have n items in common with Test C. Again, overlap with Tests A and B was not controlled.

5. For the four tests created for each of the three levels of overlap, subsamples of the 4,000 examinee population were selected using a systematic sampling procedure. That is, every jth case was selected so that samples of 100, 300, 500

1000, and 2000 were obtained. Sampling began with a different first case for each of the tests so that different examinees would be included on each of four tests for each sample size. The 2000 sample tests had substantial overlap of examinees, resulting in interdependence of the tests. Because of this the results for the 2000 sample tests should be interpreted cautiously.

The procedure described above resulted in 20 sets of response data (four tests for each of five sample sizes) for each level of overlap. Each of these data-sets was calibrated using each of the calibration programs, and for each sample size and item overlap level the calibrations of the four tests were linked using each linking procedure.

Analyses

The linked parameter estimates obtained from each of the combinations of sample size, overlap, and method were evaluated in two ways. First, correlations were obtained between the linked parameter estimates and the estimates obtained for the full 355 item test using all of the 4,000 examinees. These correlations were then tested to determine whether they were all estimates of the same correlation using a procedure set out by Snedecor and Cochran (1980). This test is performed using the following statistic:

$$\chi^2 = \sum_{i=1}^{k} (N_i - 3)z_i^2 - \left[\sum_{i=1}^{k} (N_i - 3)z_i^2\right] \sum_{i=1}^{k} (N_i - 3),$$

where k is the number of correlations, N_i is the sample size for correlation i, z_i is the normal deviate form of Correlation i obtained via Fisher's r to z transformation, and χ^2 is distributed as a chi-square with (k-1) degrees of freedom. The usual Fisher's r to z transformation was used to compare specific pairs of correlations.

The second type of analysis performed was the computation of the sum-of-squared-deviations quality of linking index suggested by Wright (1977) for use with the one-parameter logistic model. This index is based on the following equality:

$$\chi^2 = \frac{S_{BL}^2 Nn}{12},$$

where N is the sample size, and n is as previously defined. The value S_{BL}^2 represents the sum of the squared deviations of the differences between the two sets of estimates around the mean difference, or linking constant. This statistic was computed for both the one- and three-parameter b-values, although its applicability to the three-parameter model has yet to be determined.

The final type of analysis performed was the construction of scatter plots comparing the linked estimates and large sample estimates in order to check for non-linearity. When the scatter plots indicated

that there might be curvilinearity, eta coefficients were computed and compared to the correlation to determine the seriousness of the deviation from linearity.

Results

One-Parameter Logistic Model

The correlations that were obtained between the linked one-parameter logistic (1PL) difficulty parameter estimate sets obtained using the major axis linking method and the difficulty parameter estimates obtained from the calibration of the full ITED are presented in Table 1. The correlations are shown for the linkage using 5, 15, and 25 overlapping items, and for sample sizes of 100, 300, 500, 1000, and 2000. In addition, the correlations are presented for the linking results after each successive test was linked to the base test. The data were analyzed in this way to check for drift in the estimates (Rentz, 1978). In this report drift is defined as significant changes in the item parameter estimates as new tests are linked to an already existing item pool.

Table 1

Correlations of Linked One-Parameter Difficulty Estimates Obtained Using the Major Axis Method With the Large Sample Estimates for All Sample Sizes and Overlap Levels

Sample	5 Item Overlap			15 Item Overlap			25 Item Overlap		
Size	AB	ABC	ABCD	AB	ABC	ABCD	AB	ABC	ABCD
100	.949	.950	.958	.950	.948	.943	.972	.965	.969
300	.980	.981	.984	.981	.982	.983	.986	.985	.988
500	.979	.982	.988	.991	.991	.992	.992	.992	.993
1000	.989	.991	.994	.997	.997	.997	.997	.997	.998
2000	.998	.998	.998	.998	.998	.999	.999	.999	.999
No. of Items	99	134	165	85	116	144	75	98	115
χ^2			211.71**			345.93**			219.70**
df			4			4			4

** $p < .005$

Note. The χ^2 tests are reported here only for the ABCD sets of linked estimates.

The magnitude of the correlations in Table 1 indicates that the major axis linking method used in conjunction with the one-parameter logistic model works well. Comparisons of the correlations across sample sizes (summarized in Table 2) and overlap levels (shown in Table 1) yielded predictable results. For all three levels of overlap, the correlations increased significantly with increased sample size (χ^2=211.71, p<.005 for 5 item overlap, χ^2=345.93, p<.005 for the 15 item overlap; χ^2=219.70, p<.005 for 25 item overlap). (The χ^2 test was performed and reported only for the full set of four linked tests, denoted by ABCD in the table.)

The results showed that for the 1,000 and 2,000 sample sizes the correlations increased significantly with increased overlap (χ^2=21.42, p<.005 for the 1,000 sample; χ^2=11.89, p<.005 for the 2,000 sample size). For the 300 and 500 sample sizes the changes in the correlations across overlap levels were not significant. For the 100 sample a χ^2=6.54 (p<.05) was obtained, indicating significant differences, but the pattern is unclear. The correlation for the 100 sample size with 25 items overlapping (r=.969) was higher than the correlation with 15 items overlapping (r=.943) and with five items overlapping (r=.958), but the 15 item overlap correlation was not higher than the five item overlap, as might have been expected. It should be pointed out, however, that the statistically significant differences reported above may have little practical importance, since all of the reported correlations were so close to 1.0.

The next analysis performed on the correlations reported in Table 1 was a check for drift of estimates during the linking process. To test for drift correlations obtained for the AB, ABC, and ABCD sets of estimates were compared. No significant change in the magnitude of the correlations was found as the number of tests linked together increased from two to four, as is shown in Table 2, with the exception of the 1,000 sample size five item overlap level. An examination of the correlations for this case indicates that the correlations increases as the number of linked tests increased. However, comparisons of the AB correlation with the ABC correlation and the ABC correlation with the ABCD correlation were not significant. The comparison of the AB correlation with the ABCD correlation was, of course, significant.

The final analysis performed on the correlations report in Table 1 was to examine the corresponding scatter plots to determine whether there were any indications of curvilinearity. No indications of nonlinearity were found for any sample size or level of overlap.

The results for the sum-of-squared deviations quality of linking index are reported in Table 3. The values reported in the body of the table are the χ^2 values corresponding to the obtained squared deviations. In all but three cases, these χ^2 values were significant. In fact, these values were highly significant even for the conditions for which correlations of .999 were obtained, and where the standard deviations

for the parameter estimates was the same. It is clear from these results that this statistic has little relationship to the quality of the linking as defined in this study. Because of this finding no further analyses based on this statistic will be reported.

Table 2
Chi-Square Statistics for the Overlap and Drift Analysis of the Major Axis Method With the One-Parameter Logistic Model

Sample Size	Overlap	Drift		
		5 Item	15 Item	25 Item
100	6.54*	.85	.25	.51
300	2.13	.94	.19	.77
500	5.67	5.98	.40	.33
1000	21.42**	7.06*	.00	.00
2000	11.89**	.00	.00	.00
df	2	3	3	3

*$p < .05$.

**$p < .005$.

Note. The results of the overlap analyses are reported only for the ABCD sets of linked estimates.

Table 3
Quality of Linking Statistic for the One-Parameter Linking by Sample Size and Overlap

Sample Size	5 Item			15 Item			25 Item		
	AB	ABC	ABCD	AB	ABC	ABCD	AB	ABC	ABCD
100	13.0	170.5	404.7	345.3	143.4	354.4	381.2	401.3	867.1
300	8.3*	143.2	214.7	115.9	276.3	303.1	410.6	536.5	710.0
500	18.5	182.0	460.9	251.6	150.4	362.3	587.4	723.1	797.8
1000	17.5	111.2	597.1	147.7	143.8	271.3	564.7	554.7	811.6
2000	18.8	13.7*	305.3	.2*	132.4	160.2	481.1	373.0	492.7
Overlapping items	5	11	19	15	19	22	25	27	33
$\chi^2(.05)$	9.49	18.31	28.87	23.68	28.87	33.92	36.42	40.11	55.76
df	1	2	3	1	2	3	1	2	3

*These values were not significant at the .05 level. All others indicated a significant difference from a quality linking.

Three-Parameter Logistic Model

Due to the numerous procedures used with the three-parameter logistic (3PL) model the results of these analyses are relatively complex. To facilitate the presentation of the results, they will be presented for each procedure separately. Only after the specific results for each of the procedures have been presented will the results of the comparisons of the procedures be presented. It should be pointed out that in the following analyses, which contain multiple comparisons, no attempt was made to control the experimentwise error rate. The purpose of these analyses is to compare the relative qualities of the procedures, not judge them in any absolute terms.

Major Axis Method Table 4 contains the correlations between the large sample estimates and the estimates obtained from the major axis linking procedure for the five item overlap level. The correlations obtained using both the ANCILLES and LOGIST program are reported. The same data for the 15 item overlap level are shown in Table 5, while Table 6 shows the data for the 25 item overlap level. At the bottom of Tables 4 through 6 are shown the obtained chi-squares from the tests for significant changes in the correlations across sample sizes. As can be seen in these tables, the correlations increase significantly with increased sample size for both the ANCILLES and LOGIST estimates at all three levels of overlap, with the exception of the LOGIST c-values for the five item overlap level.

The tests to determine the significance of the changes in the correlations as the number of common items increased are summarized in Table 7 for both the ANCILLES and LOGIST estimates. For the ANCILLES estimates the obtained chi-squares were significant only for the a- and b-values for the 100 and 500 sample sizes. For the 100 sample size the correlations of the a- and b- values with the large sample estimates did not change significantly as overlap increased from five to 15 items. However, as the overlap increased from 15 to 25 the increase in the correlations was significant. At the 500 sample size, the a-value correlation increased significantly only between the 15 and 25 item overlap levels. The b-value correlation, however, increased significantly only between the five and 15 item overlap levels, and not between the 15 and 25 item overlap levels.

The LOGIST estimate correlations changed significantly as the number of common items increased in all cases except the c-values for the 100, 300, and 500 sample sizes. The pair-wise comparisons of the correlations (five item vs. 15 item and 15 item vs. 25 item) did not reveal any consistent pattern of change as overlap increased. For the a-values, the correlations for the 100 and 1,000 sample sizes did not increase between the five and 15 item overlap levels, but did increase significantly when the overlap was increased to 25 items. The 300 and 500 sample a-value correlations significantly decreased as overlap increased from five to 15 items, but increased when overlap increased to 25 items. For the 300 sample size the 25 item overlap a-value corre-

Table 4

Correlations of Linked Three-Parameter Estimates Obtained Using the Major Axis Method With Large Sample Estimates for the 5 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.306	.347	.327	.603	.565	.498
	b	.685	.726	.755	.501	.472	.427
	c	.382	.422	.423	.094	.072	.170
300	a	.430	.613	.593	.727	.726	.749
	b	.865	.887	.902	.712	.731	.744
	c	.623	.578	.617	.316	.252	.239
500	a	.757	.765	.770	.692	.786	.743
	b	.906	.924	.935	.730	.707	.666
	c	.697	.725	.728	.314	.237	.244
1000	a	.862	.867	.870	.828	.885	.668
	b	.945	.953	.958	.875	.868	.822
	c	.838	.838	.810	.237	.196	.208
2000	a	.898	.915	.906	.896	.933	.907
	b	.970	.976	.976	.992	.991	.987
	c	.833	.842	.828	.363	.306	.268
n_i		88	126	156	99	134	165
$\chi^2(4)$	a			137.58*			80.58*
	b			131.54*			411.21*
	c			55.85*			1.06

* $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

Table 5

Correlations of Linked Three-Parameter Estimates Obtained Using the Major Axis Method With Large Sample Estimates for the 15 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.314	.330	.421	.600	.414	.367
	b	.766	.788	.772	.762	.733	.734
	c	.493	.484	.452	-.132	.004	-.017
300	a	.698	.687	.683	.827	.511	.533
	b	.937	.924	.930	.909	.900	.907
	c	.710	.600	.566	.549	.375	.344
500	a	.664	.718	.756	.523	.520	.439
	b	.965	.959	.963	.824	.809	.829
	c	.717	.689	.742	.381	.486	.422
1000	a	.830	.870	.888	.632	.679	.630
	b	.965	.968	.969	.808	.826	.842
	c	.743	.734	.769	.329	.539	.563
2000	a	.916	.926	.923	.934	.948	.945
	b	.969	.977	.978	.987	.988	.989
	c	.769	.805	.815	.531	.676	.588
n_i		82	113	141	85	116	144
χ^2 (4)	a			119.29*			181.41*
	b			129.85*			240.91*
	c			42.01*			43.80*

* $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

Table 6

Correlations of Linked Three-Parameter Estimates Obtained Using the Major Axis Method With Large Sample Estimates for the 25 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.654	.590	.613	.521	.557	.614
	b	.774	.818	.870	.769	.772	.757
	c	.193	.310	.453	.270	.234	.084
300	a	.763	.791	.746	.849	.827	.811
	b	.904	.908	.921	.928	.921	.930
	c	.529	.554	.634	.659	.285	.249
500	a	.822	.850	.861	.801	.827	.854
	b	.961	.948	.958	.980	.978	.979
	c	.539	.553	.638	.493	.334	.326
1000	a	.889	.904	.899	.928	.918	.915
	b	.966	.957	.965	.985	.988	.984
	c	.626	.647	.698	.605	.378	.325
2000	a	.958	.944	.942	.980	.965	.972
	b	.979	.971	.978	.996	.996	.996
	c	.812	.738	.780	.822	.488	.473
n_i		70	93	109	75	98	115
$\chi^2(4)$	a			72.18*			124.31*
	b			56.63*			264.99*
	c			17.29*			10.92**

* p .005

** $p <$ .05

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

Table 7

Chi-Square Statistics for the Overlap and Drift Analyses of the Major Axis Method Using the Three-Parameter Logistic Model Using the ANCILLES and LOGIST Programs

mple ze	Parameter	Overlap		Drift					
				ANCILLES			LOGIST		
		ANCILLES	LOGIST	5 Item	15 Item	25 Item	5 Item	15 Item	25 Item
	a	8.97*	6.83*	.45	1.06	.4	1.46	5.07	.86
00	b	8.37*	25.06**	1.16	.16	4.02	.51	.24	.07
	c	.15	1.72	.15	.16	3.76	.81	1.29	2.08
	a	5.19	19.75**	3.65	.04	.59	.24	22.62**	.63
00	b	2.42	38.93**	1.54	.42	.51	.30	.15	.22
	c	.83	1.22	.33	3.25	1.32	.50	3.78	14.45**
	a	6.57*	41.77**	.05	1.84	.76	2.46	.99	1.30
00	b	6.63*	149.97**	2.10	.41	1.00	1.11	.21	.10
	c	2.70	3.02	.24	.73	1.27	.46	.86	2.04
	a	1.14	50.45**	.06	2.52	.25	26.03**	.58	.30
00	b	1.90	123.57**	1.08	.18	.77	3.03	.64	1.24
	c	4.31	14.34**	.67	.55	.80	.09	5.04	6.09
	a	4.07	25.48**	.53	.20	1.21	3.43	.72	3.46
00	b	.23	11.81**	.79	1.72	1.49	1.86	.44	.00
	c	1.17	12.26**	.15	.81	1.41	.66	2.70	22.23**
lf		2	2	3	3	3	3	3	3

*$p<.05$
**$p<.005$

lation was not significantly greater than the five item overlap correlation. But the 500 sample 25 item overlap correlation was significantly greater than the five item overlap correlation.

The pattern for the LOGIST b-values was somewhat more consistent than for the a-values. As sample size increased the change in correlation magnitude tended to occur between the 15 and 25 item overlaps. For the 100 and 300 samples the correlation increased significantly as overlap increased from five to 15 items, but did not increase significantly as overlap increased to 25 items. The 500 sample correlation increased both as the overlap increased from five to 15 items and as the overlap increased from 15 to 25 items. For the 1000 and 2000 samples the correlation did not increase as overlap increased from five to 15 items, but did increase significantly as overlap increased from 15 to 25 items.

There were no significant changes in the correlations of the LOGIST c-values for the 100, 300, and 500 samples. For the 1,000 sample the correlation increased as overlap increased from five to 15 items, but decreased as overlap increased to 25 items. The 2,000 sample c-value correlation increased between the five and 15 item overlap levels, but did not change significantly as overlap increased to 25 items.

For the ANCILLES estimates no parameter drift was found, as can be seen in Table 7. For the LOGIST estimates there was some parameter drift during linking. At the five item overlap level significant differences in the a-value correlations were found for the 1,000 sample. The a-value ABC correlation was not significantly different from the AB correlation. However, both the AB and ABC correlations were significantly higher than the ABCD correlation. For the 15 item overlap level the only significant drift was for the 300 sample a-values. In this case the AB correlation was significantly higher than both the ABC and ABCD correlations. There was no significant difference between the ABC and ABCD correlations. The only drift found for the 25 item overlap level was for the c-values at the 300, 1000, and 2000 sample sizes. In all three cases the AB correlation was significantly greater than the ABC and ABCD correlations, and in none of these cases was there a significant difference between the ABC and ABCD correlations.

Least Squares Method The correlations of the large sample estimates with the ANCILLES and LOGIST estimates linked using the least squares method are shown for the five, 15 and 25 item overlap levels in Tables 8, 9, and 10, respectively. At the bottom of these tables are summaries of the analyses to determine whether the correlations changed significantly as sample size increased. As can be seen in the tables, in all cases except for the LOGIST c-values for the five item overlap level the correlations increased significantly with increased sample size. This was true for both the ANCILLES and LOGIST estimates at each overlap level. The LOGIST c-value correlations for the five item overlap level did not change significantly with increased sample size, as was the case with the LOGIST c-values for the five item overlap level when the major axis method was used.

The analyses to determine whether the correlations for this method changed significantly as the number of overlapping items increased are summarized in Table 11 for both the ANCILLES and LOGIST estimates. The ANCILLES a-value correlations changed significantly with increased overlap for the 300, 500, and 2000 sample sizes, but not for the 100 and 1000 samples. The 300 sample a-value correlation increased significantly as overlap increased from five to 15 items, but did not change as overlap increased from 15 to 25 items. For the 500 sample the increase from 15 to 25 items resulted in a significant increase in the correlation, but the correlation did not change as overlap went from five to 15 items. The a-value correlation for the 2000 sample did not increase significantly as overlap increased from five to 15 items, nor did it increase from 15 to 25 items. The change in the correlation was significant only as overlap increased from five to 25 items.

Table 8

Correlations of Linked Three-Parameter Estimates Obtained Using the Least Squares Method With Large Sample Estimates for the 5 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.558	.486	.490	.630	.668	.707
	b	.589	.574	.543	.485	.450	.422
	c	.382	.422	.423	.094	.072	.170
300	a	.250	.463	.465	.725	.774	.779
	b	.860	.872	.893	.719	.736	.743
	c	.623	.578	.617	.316	.252	.239
500	a	.690	.745	.764	.700	.795	.839
	b	.892	.909	.924	.731	.704	.656
	c	.697	.725	.728	.314	.237	.244
1000	a	.825	.863	.870	.821	.860	.890
	b	.942	.954	.960	.876	.868	.811
	c	.838	.838	.810	.237	.196	.208
2000	a	.857	.910	.904	.886	.927	.937
	b	.969	.977	.977	.992	.991	.987
	c	.833	.842	.828	.363	.306	.268
n_i		88	126	156	99	134	165
$\chi^2(4)$	a			124.94*			69.64*
	b			232.77*			415.03*
	c			55.85*			1.05

* $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

Table 9

Correlations of Linked Three-Parameter Estimates Obtained Using the Least Squares Method With Large Sample Estimates for the 15 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.394	.369	.419	.589	.695	.713
	b	.709	.672	.670	.707	.670	.662
	c	.493	.484	.452	-.132	.004	-.017
300	a	.678	.682	.676	.859	.850	.837
	b	.950	.932	.928	.909	.888	.896
	c	.710	.600	.566	.550	.375	.344
500	a	.612	.687	.726	.777	.835	.819
	b	.960	.951	.952	.808	.789	.801
	c	.717	.689	.742	.381	.486	.422
1000	a	.864	.687	.892	.921	.925	.922
	b	.961	.964	.965	.796	.796	.806
	c	.743	.734	.769	.329	.539	.563
2000	a	.916	.926	.924	.947	.957	.954
	b	.970	.977	.978	.987	.988	.989
	c	.769	.805	.815	.531	.676	.588
n_i		82	113	141	85	116	144
$\chi^2(4)$	a			123.00***			85.49***
	b			167.79***			278.44***
	c			41.15***			43.80***

*** $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

Table 10

Correlations of Linked Three-Parameter Estimates Obtained Using the Least Squares Method With Large Sample Estimates for the 25 Item Overlap Level

Sample Size	Parameter	ANCILLES			LOGIST		
		AB	ABC	ABCD	AB	ABC	ABCD
100	a	.651	.633	.606	.715	.735	.731
	b	.813	.861	.861	.769	.752	.721
	c	.193	.310	.453	.270	.234	.084
300	a	.762	.782	.744	.896	.887	.870
	b	.910	.903	.917	.912	.900	.912
	c	.529	.554	.634	.659	.285	.249
500	a	.821	.846	.860	.838	.851	.884
	b	.960	.949	.956	.977	.973	.974
	c	.539	.553	.638	.493	.334	.326
1000	a	.886	.903	.900	.936	.928	.924
	b	.967	.958	.965	.986	.989	.986
	c	.626	.647	.698	.605	.378	.325
2000	a	.959	.954	.950	.982	.965	.973
	b	.979	.971	.977	.996	.996	.997
	c	.812	.738	.780	.822	.488	.473
n_i		70	93	109	75	98	115
$\chi^2(4)$	a			81.94**			89.52**
	b			57.89**			298.62**
	c			17.28**			10.94*

* $p < .05$.

** $p < .005$.

Note. The row labeled n_i indicates the number of items used to compute the correlations. The ANCILLES program has fewer values because it automatically deletes poor items.

The ANCILLES b-value correlation increased with increased overlap only for the 500 sample size. The correlation for the 500 sample increased significantly as overlap increased from five to 15 items, but not when overlap increased from 15 to 25 items. The ANCILLES c-value correlation did not change significantly with increased overlap for any sample size.

The LOGIST a-value correlation increased with increased overlap only for the 2000 sample. The increase in the 2000 sample a-value correlation was significant as overlap increased from 15 to 25 items, but not as overlap increased from five to 15 items.

The LOGIST b-value correlation increased significantly with increased overlap for all sample sizes. For the 100 and 300 samples the b-value correlation increased significantly as the overlap increased from five to 15 items, but not as overlap increased from 15 to 25 items. For the 500 sample the increase was significant as overlap increased from five to 15 items as well as from 15 to 25 items. The LOGIST b-value correlation for the 1000 and 2000 samples increased significantly between the 15 and 25 item overlap levels, but not as overlap increased from five to 15 items.

The LOGIST c-value correlation changed as overlap increased only for the 1000 and 2000 samples. The 1000 sample c-value correlation increased as the overlap increased from five to 15 items, but decreased as the overlap increased from 15 to 25 items. The 2,000 sample c-value correlation increased as overlap increased from five to 15 items but did not change as overlap increased from 15 to 25 items.

The analyses to detect drift are also summarized in Table 11. As can be seen, no significant drift was found for the ANCILLES estimates. Drift in the LOGIST estimates was found for four cases. The first case was the 500 sample a-value correlation for the five item overlap level. For this case the correlation did not change significantly when the number of linked tests increased from two to three, nor when the number of linked tests went from three to four. However, the correlation for the two test set was significantly lower than the correlation for the set of four tests. No cases of drift were detected for the LOGIST estimates for the 15 item overlap level, but there were three cases of significant drift for the 25 item overlap level. For the 25 item overlap 300 sample c-values there was a significant decrease in the correlation when the third test was linked to the first two. There was no significant change in correlation when the fourth test was added. The same pattern occurred for the c-value correlation for the 1000 and 2000 sample sizes.

Least Squares Method With Outlier Deletion The results for the least squares method with outlier deletion were very similar to the results for the least squares method and therefore will not be discussed in great detail. The correlations with the large sample estimates obtained for the estimates yielded by the outlier deletion linking procedure for the five, 15, and 25 item overlap levels are shown in Tables 12 through 14, respectively. (Note that this procedure was used only in conjunction with the LOGIST program.) The analyses of the effects of increased sample size are summarized at the bottom of Tables 12 through 14. As can be seen, the correlations increased with increased sample size in all cases except the five item overlap c-values, which has been a consistent finding when the LOGIST c-values have been used.

Table 11

Chi-Square Statistics for the Overlap and Drift Analyses of the Least Squares Method with the Three-Parameter Logistic Model Using the ANCILLES and LOGIST Programs

Sample Size	Parameter	Overlap		Drift					
				ANCILLES			LOGIST		
		ANCILLES	LOGIST	5 Item	15 Item	25 Item	5 Item	15 Item	25 Item
100	a	3.78	.24	.59	.22	.28	1.03	2.67	.07
	b	30.42**	16.45**	.25	.32	1.28	.39	.36	.57
	c	.15	2.45	.16	.16	3.65	.81	.97	2.14
300	a	14.81**	5.82	3.90	.01	.39	.97	.36	.70
	b	3.16	28.23**	1.30	1.93	.31	.18	.58	29
	c	.82	1.22	.33	3.26	1.29	.50	3.78	14.31**
500	a	8.30*	3.89	2.21	2.22	.69	7.39*	1.24	1.73
	b	6.59*	132.89**	1.83	.54	.69	1.38	.13	.27
	c	2.62	3.02	.26	.68	1.32	.46	.86	2.02
1000	a	1.39	3.45	1.48	.77	.33	4.17	.04	.36
	b	.46	152.82**	1.97	.18	.78	4.23	.05	.98
	c	4.40	14.34**	.56	.43	.81	.09	5.04	6.06*
2000	a	7.05*	12.83**	3.54	.21	.42	5.93	.60	4.55
	b	.08	15.60**	1.46	1.34	1.25	4.83	.44	.00
	c	1.16	12.26**	.17	.74	1.34	.66	2.70	21.71**
df		2	2	3	3	3	3	3	3

* $p<.05$
** $p<.005$

A summary of the overlap analyses for the outlier deletion procedure is shown in Table 15. The a-value correlation with the large sample estimates increased with increased overlap only for the 300 and 2000 sample sizes. The 300 sample a-value correlation increased significantly from the five to 25 item overlap levels, but not between the five and 15 item overlap levels nor between the 15 and 25 item overlap levels. The 2000 sample a-value correlation increased significantly as overlap increased from 15 to 25 items, but not as overlap increased from five to 15 items.

Table 12

Correlations of Linked Three-Parameter Estimates Obtained Using the Maximum Likelihood Method and the Least Squares Method With Outlier Deletion With Large Sample Estimates for the 5 Item Overlap Level

Sample Size	Parameter	Maximum Likelihood	Least Squares With Outlier Deletion		
			AB	ABC	ABCD
100	a	.685	.631	.674	.709
	b	.476	.485	.448	.420
	c	.139	.094	.072	.170
300	a	.816	.725	.769	.774
	b	.788	.719	.729	.728
	c	.273	.316	.252	.239
500	a	.826	.700	.795	.838
	b	.665	.731	.706	.655
	c	.206	.314	.237	.244
1000	a	.907	.821	.873	.903
	b	.831	.876	.868	.809
	c	.196	.237	.196	.208
2000	a	.948	.886	.926	.936
	b	.988	.992	.991	.992
	c	.353	.363	.306	.268
n_i		165	99	134	165
$\chi^2(4)$	a	89.97*			72.40*
	b	407.17*			541.71*
	c	5.02			1.05

Note. The row labeled n_i indicates the number of items used to compute the correlations.

Table 13

Correlations of Linked Three-Parameter Estimates Obtained Using the Maximum Likelihood Method and the Least Squares Method With Outlier Deletion With Large Sample Estimates for the 15 Item Overlap Level

Sample Size	Parameter	Maximum Likelihood	Least Squares With Outlier Deletion		
			AB	ABC	ABCD
100	a	.735	.585	.692	.711
	b	.673	.693	.633	.633
	c	.096	-.132	.004	-.017
300	a	.793	.858	.850	.836
	b	.943	.908	.889	.895
	c	.262	.550	.375	.344
500	a	.811	.778	.835	.818
	b	.923	.809	.786	.773
	c	.269	.381	.486	.422
1000	a	.917	.921	.925	.923
	b	.951	.796	.810	.800
	c	.400	.329	.539	.563
2000	a	.945	.947	.960	.954
	b	.983	.987	.988	.989
	c	.407	.531	.676	.588
$\chi^2(4)$	a	73.16**			86.00**
	b	180.97**			297.13**
	c	10.65**			10.65*

* $p < .05$

** $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations.

Table 14

Correlations of Linked Three-Parameter Estimates Obtained Using the Maximum Likelihood Method and the Least Squares Method With Outlier Deletion With Large Sample Estimates for the 25 Item Overlap Level

Sample Size	Parameter	Maximum Likelihood	Least Squares With Outlier Deletion		
			AB	ABC	ABCD
100	a	.673	.718	.730	.734
	b	.634	.769	.770	.738
	c	.063	.270	.234	.084
300	a	.857	.896	.887	.870
	b	.940	.908	.913	.923
	c	.388	.659	.285	.249
500	a	.879	.838	.855	.886
	b	.987	.976	.969	.968
	c	.296	.493	.334	.326
1000	a	.945	.936	.928	.924
	b	.992	.994	.994	.990
	c	.407	.605	.378	.325
2000	a	.974	.982	.965	.973
	b	.997	.996	.996	.996
	c	.445	.822	.488	.473
n_i		115	75	98	115
$\chi^2(4)$	a	118.76**			88.19**
	b	381.23**			302.55**
	c	12.56*			10.94*

* $p < .05$

** $p < .005$

Note. The row labeled n_i indicates the number of items used to compute the correlations.

The b-value correlation increased significantly with increased overlap at all sample sizes. The 100 and 300 sample size b-value correlations increased significantly as overlap increased from five to 15 items, but not as overlap increased from 15 to 25 items. For the 500 sample the increase in correlation was significant as overlap increased from five to 15 items as well as from 15 to 25 items. The 1000 and 2000 sample b-value correlations increased significantly as overlap increased from 15 to 25 items, but not as overlap increased from five to 15 items.

For the c-values the correlation changed significantly with increased overlap only for the 1000 and 2000 sample sizes. The 1000 sample correlation increased significantly as overlap increased from five to 15 items and decreased as overlap increased to 25 items. The 2000 sample c-value correlation increased as overlap increased from five to 15 items, but did not change as overlap increased to 25 items.

Table 15

Chi-Square Statistics for the Overlap Analyses of the Maximum Likelihood Method and the Least Squares Method With Outlier Deletion and the Drift Analyses for the Least Squares Method with Outlier Deletion

Sample Size	Parameter	Overlap		Drift (Outlier Deletion)		
		Max. Like.	Outlier Deletion	5 Item	15 Item	25 Item
100	a	.05	.23	1.28	2.64	.06
	b	7.26*	17.11**	.39	.58	.38
	c	.92	2.45	.81	.97	2.17
300	a	2.55	6.27*	.81	.36	.70
	b	46.91**	36.43**	.03	.53	.44
	c	2.86	1.22	.50	3.78	14.45**
500	a	3.97	4.16	6.97*	1.24	1.74
	b	199.17**	114.75**	1.41	.42	1.08
	c	5.57	3.02	.46	.86	2.04
1000	a	5.06	1.48	6.61*	.03	.35
	b	169.52**	198.86**	4.64	.09	4.89
	c	16.66**	14.34**	.09	5.04	6.09*
2000	a	11.51**	13.12**	5.75	1.07	4.75
	b	24.03**	9.02*	.43	.44	.00
	c	8.36*	12.26**	.66	2.70	21.96**
df		2	2	3	3	3

*$p < .05$
**$p < .005$

A summary of the drift analyses for the outlier deletion method also appears in Table 15. No significant drift was found for the 15 item overlap level, but several instances of drift were found for the five and 25 item overlap levels. For the five item overlap level there was significant drift for the a-value correlations for the 500 and 1000 sample sizes. In both cases there was no significant change in the correlation as the number of linked tests increased from two to three and from three to four. The only significant difference was between the AB and ABCD correlations. For the 25 item overlap level significant drift was found for the c-value correlation for the 300, 1000, and 2000 sample sizes. In all three cases the c-value correlation decreased significantly as the third test was linked, and in none of these three cases did the correlation change when the fourth test was added.

Maximum Likelihood Method Tables 12 through 14 show the correlations obtained between the large sample estimates and the estimates yielded by the maximum likelihood linking method for the five, 15, and 25 item overlap levels, respectively. The summary of the sample size effect analyses shown at the bottom of these tables indicates that the correlations increased significantly with increased sample size in all cases but the five item overlap c-values.

Table 15 shows the results of the overlap analyses for the maximum likelihood method. As can be seen in the table, the results of the overlap analyses for this method are similar to the results of the overlap analyses for the least squares method. The a-value correlation increased as overlap increased only for the 2,000 sample size. The 2,000 sample a-value correlation did not change as overlap increased from five to 15 items, but increased significantly as overlap increased from 15 to 25 items.

The b-value correlation increased with increased overlap at all sample sizes. At the 100 and 300 sample sizes the b-value correlation increased as overlap increased from five to 15 items, but not as overlap went from 15 to 25 items. For the 500 and 1,000 samples the increase in correlation was significant as overlap went from five to 15 items as well as when overlap increased from 15 to 25 items. The increase in correlation for the 2,000 sample was significant as overlap increased from 15 to 25 items, but not as overlap went from five to 15 items.

The c-value correlation increased with increased overlap only for the 1,000 and 2,000 samples. In both cases the c-value correlation increased as overlap increased from five to 15 items, but did not change significantly as overlap increased from 15 to 25 items.

The maximum likelihood method involves the simultaneous calibration of the tests to be linked. Therefore, no intermediate estimates (AB and ABC sets of estimates) are obtained for this linking method. As a result, there could be no drift analyses for this procedure.

The examination of the scatter plots for the three-parameter correlations indicated some cases for which there appeared to be deviations from linearity. However, for none of the cases were the eta coefficients that were computed different from the correlations, indicating that there were no significant deviations from linearity. This was true for all sample sizes, all levels of overlap, and for every linking procedure.

Comparison of the Procedures

Table 16 summarizes the correlations obtained using the seven linking procedures for the five item overlap level. In each row the highest correlation is underlined, with multiple underlining indicating no significant differences among the underlined correlations. The one-parameter logistic model results for the five item overlap level are reported in Table 16, but were not considered when the highest correlations were underlined. The one-parameter correlations were presented for comparison purposes. The one-parameter results were presented previously and were shown in Table 1.

One result that can be seen immediately from Table 16 is that for the five item overlap level the major axis linking method with LOGIST estimates yielded lower correlations overall than the other procedures. However, when ANCILLES estimates were used the major axis method yielded correlations that were in most cases as high or higher than the correlations obtained for the other procedures. In 11 of 15 cases the correlations for the major axis method using ANCILLES estimates were underlined, a total which was higher than any of the other procedures for the five item overlap level.

The next highest total was for the least squares method using ANCILLES estimates, which had 10 correlations underlined. The remaining procedures, least squares using LOGIST estimates, outlier deletion using LOGIST estimates, and the maximum likelihood method using LOGIST, were much the same. When LOGIST estimates were used, the least squares method had five correlations underlined, as did the maximum likelihood method. The outlier deletion method had six correlations underlined.

There is a relatively clear pattern evident in Table 16. The correlations obtained using LOGIST estimates with the least squares, outlier deletion, and maximum likelihood methods were almost identical. Only for the 2,000 sample size $\underline{b}$-value correlations was there a difference. The correlation obtained for the outlier deletion method for the 2,000 sample size $\underline{b}$-values was the highest of all the methods. With the exception of the major axis method, it appears that for the five item overlap level the methods were of about the same quality when LOGIST estimates were used. It also appears that across the sample sizes the best results for the $\underline{a}$-value linking were obtained using LOGIST estimates. For the 300 and 500 sample sizes the correlations obtained for the $\underline{a}$-values using ANCILLES estimates were as high as those for the LOGIST estimates, but for the other three sample sizes, the 100, 1000, and 2000 sample sizes, the correlations were higher when LOGIST estimates were used.

Table 16

Correlations of Linked Estimates With Large Sample Estimates For the One- and Three-Parameter Linking Procedures for the 5 Item Overlap Level

Sample Size	Parameter	1PL	3PL Estimates					
			ANCILLES		LOGIST			
			Major Axis	Least Squares	Major Axis	Least Squares	Outlier Deletion	Max. Like.
100	a	.958	.327	.490	.498	.707	.709	.685
	b		.755	.543	.427	.422	.420	.476
	c		.423	.423	.170	.170	.170	.139
300	a	.984	.593	.465	.749	.779	.774	.816
	b		.902	.893	.744	.743	.728	.788
	c		.617	.617	.239	.239	.239	.273
500	a	.988	.770	.764	.743	.839	.838	.826
	b		.935	.924	.666	.656	.655	.665
	c		.728	.728	.244	.244	.244	.206
1000	a	.994	.870	.870	.668	.890	.903	.907
	b		.958	.960	.822	.811	.809	.831
	c		.810	.810	.208	.208	.208	.196
2000	a	.998	.906	.904	.907	.937	.936	.948
	b		.976	.977	.987	.987	.992	.988
	c		.828	.828	.268	.268	.268	.353

Note. The largest correlation in each row is underlined for the 3PL procedures. More than one underlined value per row indicates no significant difference between the underlined values.

When ANCILLES estimates were used there was little difference between the major axis and least squares methods. The major axis method correlation was higher for the 100 sample size b-values, but for all other sample sizes the correlations were equally high.

Overall, for the five item overlap level it appears that the major axis method using ANCILLES estimates yields the best results. However, it is clear that for a-value linking the LOGIST estimates using any of the procedures except the major axis method are best. For the b-value linking the ANCILLES estimates with the major axis method yield the best results except for the 2000 sample size, in which case outlier deletion with LOGIST estimates appears to be best. For the c-values, which are not linked, the correlations were markedly higher when the ANCILLES estimates were used.

For all sample sizes the b-value correlations were higher for the one-parameter linking than any of the other procedures If no estimates are needed for item discrimination or guessing, the one-parameter linking is clearly superior to three-parameter linking when judged in terms of correlations with large sample estimates.

Table 17 summarizes the correlations obtained from all of the procedures for the 15 item overlap level. Again, the highest correlations in each row are underlined, with multiple underlining indicating no significant differences among the underlined values.

Table 17

Correlations of Linked Estimates With Large Sample Estimates For the One- and Three-Parameter Linking Procedures for the 15 Item Overlap Level

Sample Size	Parameter	1PL	3PL Estimates					
			ANCILLES		LOGIST			
			Major Axis	Least Squares	Major Axis	Least Squares	Outlier Deletion	Max. Like.
100	a	.943	.421	.419	.367	.713	.711	.735
	b		.772	.670	.734	.662	.633	.673
	c		.452	.452	-.017	-.017	-.017	.096
300	a	.983	.683	.676	.533	.837	.836	.793
	b		.930	.928	.907	.896	.895	.943
	c		.566	.566	.344	.344	.344	.262
500	a	.992	.756	.726	.439	.819	.818	.811
	b		.963	.952	.829	.801	.773	.923
	c		.742	.742	.422	.422	.422	.269
1000	a	.997	.888	.892	.630	.922	.923	.917
	b		.969	.965	.842	.806	.800	.951
	c		.769	.769	.563	.563	.563	.400
2000	a	.999	.923	.924	.945	.954	.954	.945
	b		.978	.978	.989	.989	.989	.983
	c		.815	.815	.588	.588	.588	.407

Note. The largest correlation in each row is underlined for the 3PL procedures. More than one underlined value per row indicates no significant difference between the underlined values.

The pattern of correlations in Table 17 is much like what was obtained for the five item overlap level. One difference that can be seen is that the correlations in most cases were higher for the smaller sample sizes than for the five item overlap level, as was indicated by the overlap analyses previously discussed. For those correlations obtained using ANCILLES estimates, there were only two other differences. The correlation obtained for the 100 sample b-values for the 15 item overlap level using the least squares method with ANCILLES estimates was not significantly lower than the major axis method correlation, while it was lower than the major axis method correlation for the five item overlap. The opposite was true for the 500 sample a-value correlation. For the 15 item overlap level 500 sample a-values, the correlation obtained for the least squares method using ANCILLES estimates was significantly lower than the major axis method correlation using ANCILLES estimates, while for the five item overlap the two correlations were not significantly different.

There were several changes in the correlations obtained for the procedures using LOGIST estimates when overlap increased from five to 15 items. One change was that for the 15 item overlap level the major axis method using LOGIST estimates yielded correlations for the 2,000 sample size that were not significantly lower than the correlations for the other procedures, as was the case for the five item overlap level. The use of LOGIST estimates with the other procedures still appeared to be superior to use of the ANCILLES estimates when a-values were linked, although the ANCILLES estimates gave equally high a-value correlations for the 1,000 sample size and for the 500 sample size when the major axis method was used. The major axis method using LOGIST estimates was still clearly inferior to the other procedures, and there were still few differences between the other procedures when LOGIST estimates were used. There were no differences in the procedures for the a-values. However, for the b-values the maximum likelihood procedure was superior to the other procedures using LOGIST estimates for the 300, 500, and 1,000 samples. For the 2,000 sample b-values the maximum likelihood procedure yielded a lower correlation than the other procedures when LOGIST estimates were used. One other interesting result appears in Table 17. The outlier deletion correlation for the 100 sample b-values is the only correlation that is significantly lower than the others.

Overall, the results for the 15 item overlap level are not much different than for the five item overlap level. The major axis method using ANCILLES estimates yielded higher overall correlations than the other procedures, although the LOGIST estimates yielded correlations for the 15 item overlap level that compared more favorably with the correlations for the ANCILLES estimates than was the case with the five item overlap level. For a-value linking the procedures using LOGIST estimates, with the exception of the major axis method, appeared to be superior. At all sample sizes except the 2,000 sample size the procedures using ANCILLES estimates appeared to give better overall results for b-value linking, and in all cases the correlations were higher for the c-values when ANCILLES estimates were used. Again, c-values were not linked.

Table 18 summarizes the results for the 25 item overlap level. The results are somewhat different from the results for the five and 15 item overlap levels. One important result reported in Table 18 is that, with the exception of the c-values and 100 sample size b-values, the maximum likelihood procedure appears to be the procedure of choice. For the a- and b-values the maximum likelihood procedure correlations were as high or higher than the correlations for any procedure for all sample sizes, except for the 100 sample b-values. For the c-values the ANCILLES estimates yielded higher correlations, as they did for the 100 sample b-values.

Table 18

Correlations of Linked Estimates With Large Sample Estimates For the One- and Three-Parameter Linking Procedures for the 25 Item Overlap Level

Sample Size	Parameter	1PL	3PL Estimates					
			ANCILLES		LOGIST			
			Major Axis	Least Squares	Major Axis	Least Squares	Outlier Deletion	Max. Like.
100	a	.969	.613	.606	.614	.731	.734	.673
	b		.870	.861	.757	.721	.738	.634
	c		.453	.453	.084	.084	.084	.063
300	a	.988	.746	.744	.811	.870	.870	.857
	b		.921	.917	.930	.912	.923	.940
	c		.634	.634	.249	.249	.249	.388
500	a	.993	.861	.860	.854	.884	.886	.879
	b		.958	.956	.979	.974	.968	.987
	c		.638	.638	.326	.326	.326	.296
1000	a	.998	.899	.900	.915	.924	.924	945
	b		.965	.965	.984	.986	.990	.992
	c		.698	.698	.325	.325	.325	.407
2000	a	.999	.942	.950	.972	.973	.973	.974
	b		.978	.977	.996	.997	.996	.997
	c		.780	.780	.473	.473	.473	.445

Note. The largest correlation in each row is underlined for the 3PL procedures. More than one underlined value per row indicates no significant difference between the underlined values.

After the maximum likelihood procedure, the next best procedure for linking b-values appeared to be the outlier deletion procedure, followed closely by the major axis method using ANCILLES estimates and the least squares method using ANCILLES estimates. For a-value linking the other three procedures using LOGIST estimates appeared to be as good as the maximum likelihood procedure. Unlike the five and 15 item overlap levels, the major axis method using LOGIST estimates did not appear to be inferior to the other linking procedures.

At the 25 item overlap level, as was the case with the other overlap levels, the one-parameter linking of b-values was superior to the three-parameter procedures, especially for the smaller sample sizes. A sample size of 500 was required for the three-parameter b-value correlations to exceed the one-parameter 100 sample size b-value correlation. When ANCILLES estimates were used the one-parameter 100 sample b-value correlation was not exceeded until the 2,000 sample size.

DISCUSSION

The purpose of this study was to investigate the properties of various procedures available for linking item parameter estimates for the one- and three- parameter logistic models. The properties of interest included: (a) the sample size requirements of the procedures; (b) the overlap requirements of the procedures; (c) the degree of drift when new tests are linked to an existing pool; and (d) the relative quality of the procedures. These properties were investigated by obtaining and analyzing the correlations between the linked item parameter estimates yielded by the linking procedures and item parameter estimates yielded by a large sample calibration of the items. Before discussing the results of those analyses, the use of correlations as criteria for judging the quality of linking will be discussed.

For the 3PL model, the use of correlations as criteria for judging the quality of the linking of b-values appears to be reasonable. For the 1PL model a slope of one is also required. Two sets of b-values for the same items should be linearly related, and any departure from linearity that occurs will be reflected in the correlations. Moreover, as long as the correlation is one, it is not necessary to consider further conditions, such as the intercept of the regression line. For a- and c-values, however, not only is it necessary for the correlation to be one, but the intercept of the regression line must be zero. Therefore, in the case of a-value and c-value linking, a correlation of one is a necessary but not sufficient condition. When evaluating linking procedures, then, a high correlation is not enough for concluding that the linking results are adequate.

Two other considerations must be made when using correlations as criteria for judging the adequacy of linking. First, the actual concern in judging the quality of linking is that the resulting ICC be correct, so all three parameters must be considered together. It is not enough to have good linking of the estimates of one parameter (unless, of course, a one-parameter model is being used). Second, it is very difficult to determine how high a correlation should be in order to judge the linking to be adequate. Ideally the correlations should be one,

but in practice correlations of one are rare. In this study, correlations that were less than .9 were judged inadequate, and for the b-values still higher correlations were required for a judgement of adequacy.

Keeping in mind the considerations just set out, the results of the analyses will now be discussed. The results for the major axis procedure using the 1PL model will be discussed first. Then each procedure using the 3PL model will be discussed in terms of the sample size requirements at each overlap level. After discussing the sample size requirements of the procedures, the relative quality of the procedures will be evaluated. Also, the results obtained for the ANCILLES and LOGIST estimates will be compared. Finally, the results obtained using the 1PL and 3PL models will be compared.

In the discussion of the results of this study the c-values will not be discussed in any detail. The c-values are simply averaged during linking, so much of the differences in c-values that occurred were due to differences in the way the different calibration programs handled the c-values. For instance, the LOGIST program places rather restrictive controls on the c-values, and as a result the c-values take on a quite restricted range of values. The restriction in range resulted in low correlations that are not truly reflective of the quality of the estimates. Therefore, comparisons of c-value correlations would not be very meaningful and will not be undertaken.

One-Parameter Logistic Model

Sample Size

The comparisons of the results obtained for the 1PL model across sample sizes indicated that for all levels of overlap the correlations increased significantly with increased sample size. However, even for the 100 sample size the correlations for the 1PL model were quite high. The increase in the correlations with increased sample size was statisfically significant, but perhaps of little practical importance. A sample size of 100 appeared to be sufficient for adequate linking regardless of the level of overlap.

Overlap

For the 1000 and 2000 sample cases the correlations increased significantly with increased overlap. The smaller sample size correlations remained fairly stable as overlap increased. Based on this finding and the results of the 1PL sample size analyses, the major axis linking of the 1PL estimates appeared to be adequate for sample sizes as small as 100 and for levels of overlap as low as five. However, in a previous study (Reckase, 1977) it was found that a sample size of 300-400 was necessary for accurate estimation of the 1PL parameters. A sample of 400 with an overlap of 10 items was recommended by Wright (1977).

Drift

The results of the drift analyses indicated no significant drift for any case except the 1000 sample case for the five item overlap level.

The indication is that, although the five item overlap level appeared to be adequate, there might be some danger of drift at that overlap level. However, the problem of drift for the 1PL model was minimal even at the five item overlap level.

Quality of Linking Index

The quality of linking index for the 1PL model appeared to have little relationship to the actual quality of linking. Significant chi-squares were obtained even for the conditions for which correlations of .999 were obtained. Therefore, this index was discarded and not considered further.

Three-Parameter Logistic Model

The results obtained for the 3PL model were not consistent across the three item parameters that were estimated, nor were they the same for the two estimation procedures. Therefore, the results for the 3PL model will be discussed for one item parameter at a time. For each procedure the results will be discussed for each level of overlap as well as for both the ANCILLES and LOGIST estimate. Also, for each procedure the drift that occurred in the estimates will be discussed.

In the analysis of the sample size requirements for the procedures using the 3PL model, one factor that was considered was the stability of the correlations. However, it should be pointed out that the stability of the correlations was used to determine whether larger sample sizes would improve the correlations. Stability was not an indication of the quality of linking, since a correlation could easily become stable at a low value.

Discrimination Values

Major Axis Method For the five item overlap level the correlations for major axis linking of the a-values using LOGIST estimates never attained stability. The correlations increased as sample size increased from 100 to 300, but did not significantly change as sample size increased from 300 to 500. When sample size increased to 1000 the correlation actually decreased, and then jumped dramatically when sample size was increased to 2000. These results seem to indicate that for overlap levels as low as five, a minimum sample size of 2000 is necessary for major axis linking of LOGIST a-values.

When ANCILLES a-values were used for major axis linking at the five item overlap level the results were not much different. The correlations never decreased as sample size increased, as was the case with the LOGIST estimates, but they also never stabilized at a single value. As was the case with the major axis linking of LOGIST a-values, the linking of ANCILLES a-values using the major axis method seemed to require a minimum sample size of 2000.

When overlap was increased to 15 items, the correlations reported for the major axis linking of the ANCILLES a-values increased at every sample size, but even with the increased number of common items the minimum sample size required was still 2000. When LOGIST a-values were used the correlations reported for the major axis method did not increase as over-

lap increased from five to 15, except for the 2000 sample size case. Again, a sample size of 2000 seemed to be required for the major axis method.

When overlap was increased to 15 items, the correlations reported for the major axis linking of the ANCILLES a-values increased at every sample size, but even with the increased number of common items the minimum sample size required was still 2000. When LOGIST a-values were used the correlations reported for the major axis method did not increase as overlap increased from five to 15, except for the 2000 sample size case. Again, a sample size of 2000 appeared to be required for the major axis method.

For the 25 item overlap level the correlations reported for the major axis method were higher than for the 15 item overlap level. However, as sample size increased the correlations continued to increase for both the ANCILLES and LOGIST a-values. Once again the results indicated that a sample size of 2000 should be used for major axis linking. However, it should be pointed out that when there were 25 common items the correlations obtained for the 1000 sample size were about the same magnitude as the correlations obtained for the 2000 sample size at the five item overlap level, indicating that use of an overlap level as great as 25 items cut in half the sample size required to obtain a quality of linking equal to the quality of linking at the five item overlap level.

There were few instances of significant drift encountered during the linking of the a-values using the major axis method. When ANCILLES a-values were used there were no instances of drift. When LOGIST a-values were used drift occurred for the 1000 sample size five item overlap case and the 300 sample size 15 item overlap case. No drift was found for the 25 item overlap level. Thus, as overlap increased, drift occurred for increasingly smaller sample sizes, and with the 25 item overlap level did not occur at all.

Least Squares Method In most cases the use of outlier deletion did not significantly alter the results of the least squares linking of the LOGIST a-values. Therefore, in the discussion of the results no distinction will be made between the least squares and least squares with outlier deletion methods except in those few instances where there was a difference. For the linking of a-values there were no significant differences between those two methods.

At every overlap level, for every sample size, the correlations obtained for the least squares linking of LOGIST a-values were as high or higher than the correlations obtained using ANCILLES a-values. At no sample size did either set of correlations level off. Rather, they continued to increase with increased sample, indicating that the best sample size for this procedure was 2000, regardless of overlap or whether ANCILLES or LOGIST a-values were used. For all three of the overlap levels the least squares method using LOGIST a-values yielded correlations much higher than those obtained using ANCILLES a-values when sample sizes of 100, 300, and 2000 were used. For the 500 and 1000 sample

sizes there wasn't much difference between the correlations yielded by the two sets of estimates. For the five and 15 item overlap levels the LOGIST estimates yielded correlations for the 1000 sample size that were comparable to the correlations yielded by the ANCILLES estimates for the 2000 sample size. It appears, then, that for least squares linking of a-values using large samples LOGIST a-value estimates yielded better results than ANCILLES a-values.

As was the case with the major-axis method, for the least squares linking of ANCILLES a-values there were no incidents of drift. When LOGIST a-values were used the only drift that occurred was for the 500 sample size 15 item overlap level. When outliers were deleted, drift occurred for the 500 sample size five item overlap case and for the 1000 sample size five item overlap case.

Maximum Likelihood Method For all three overlap levels the correlations obtained for the maximum likelihood method increased with increased sample size. However, for the 25 item overlap level the correlation for the 1000 sample size was as high as the 2000 sample size correlation for the five and 15 item overlap levels, indicating perhaps that for overlap levels as high as 25 a sample size of 1000 is adequate for the maximum likelihood linking procedure. Drift was not a consideration for the maximum likelihood method, due to the simultaneous calibration of the tests.

Difficulty Values

Major Axis Method The sample size needed to obtain stable correlations for major axis linking of b-values was smaller than the sample size required for the a-values except for the 100, 300, and 500 sample size cases for the five item overlap level. For the five item overlap level the correlations for the major axis method using ANCILLES estimates became relatively stable when the sample size reached 300, although some improvement in the correlations did occur as sample size increased beyond 300. When LOGIST estimates were used the correlations did not become stable, indicating that when LOGIST estimates were used for major axis linking of b-values a sample size of 2000 or more seemed to be required.

The correlations obtained for the major axis procedure for the 15 item overlap level were higher than for the five item overlap level, but the sample size requirements were the same. A sample size of 300 still seemed to be required when ANCILLES b-values were used, while the requirement when LOGIST b-values were used was 2000.

For the 25 item overlap level the results for the ANCILLES b-values were much the same as for the 15 item overlap, with a sample size requirement of 300 indicated. For the LOGIST b-values the results were different than the 15 item overlap results. For the LOGIST b-values at the 15 item overlap level a sample size of 2000 was needed, but for the 25 item overlap a sample size of only 300 was required. For the major-axis linking of b-values there was no drift.

Least Squares Method As was the case with the a-values, for the b-values there was no practical difference between the least squares method and the least squares method with outlier deletion. Therefore, in this discussion there will be no distinction made between these two procedures. References to the least squares method using LOGIST estimates will refer to both the least squares and the least squares with outlier deletion methods.

The correlations for the five item overlap level obtained for the least squares method using ANCILLES b-values become relatively stable when samples as great as 300 were used. When LOGIST b-values were used the correlations obtained for the five item overlap level did not stabilize. The results using the LOGIST b-values indicated that a sample size of 2000 was needed.

The results for the least squares linking of b-values with 15 common items were much the same as the results for the five item overlap level. For ANCILLES b-values 300 cases seemed sufficient, while for the LOGIST b-values a sample size of 2000 was required.

For the 25 item overlap level the results for the ANCILLES b-values were the same as for the other overlap levels. A sample size of 300 appeared to be adequate. A sample size of 300 also appeared to be adequate for the LOGIST b-values. For the least squares linking of b-values there was no drift, nor was there drift when outliers were deleted.

Maximum Likelihood Method For the maximum likelihood linking of LOGIST b-values the correlations obtained for the five item overlap level did not stabilize, and a sample size requirement of 2000 was indicated. When overlap increased to 15 items the correlations became relatively stable when sample size increased beyond 100. A sample size of 300 was adequate for this procedure for the 15 item overlap level. The results were the same for the 25 item overlap level, and a sample size requirement of 300 was again indicated. As was stated previously, drift was not a consideration when linking was performed using the maximum likelihood method.

Comparisons of the 3PL Methods

The comparisons of the 3PL methods will be made in the following manner. First, the combination of sample size, overlap, calibration procedure, and linking procedure that produced the best results will be selected for each parameter. Then the best combination for small samples and low overlap levels will be selected for each parameter.

Discrimination For the linking of a-values the best combination was the 25 item overlap level for the 2000 sample case using the LOGIST a-values with any of the methods. That is, with 2000 cases and 25 items in common, it didn't matter which method was used as long as LOGIST a-values were used.

For lower levels of overlap the results were the same, except that at the five item overlap level the major axis method was not as good as the others. For small sample sizes (100 to 300) the pattern was the same with the exception that the 25 item overlap 100 sample case using the maximum likelihood method was not as good as the others. Of course, the actual correlations obtained for the 100 and 300 sample sizes were much smaller than those obtained at the 2000 sample size and appeared to be inadequate.

Difficulty For the linking of b-values the best combination was the same as for the a-values, which was any of the methods using LOGIST b-values with a sample of 2000 and 25 items in common.

For the 15 item overlap level the best combination was the same as for the 25 item overlap, except that the maximum likelihood method was not as good as the others. For the five item overlap level the best combination was the 2000 sample case for the least squares method with outlier deletion using LOGIST b-values.

For small sample sizes (100-300) the best results were obtained using the major axis and least squares methods with either the ANCILLES or LOGIST b-value estimates and 25 items in common. Again, the 100 and 300 sample size correlations were smaller than those obtained for the 2000 sample size and were probably inadequate.

Comparison of ANCILLES and LOGIST Estimates

This study was not designed to compare these two calibration procedures per se, but was to compare the linked parameter estimates obtained using these two procedures. Of course, the quality of the linking does depend on the quality of the parameter estimates.

It is clear from the results of this study that for the linking of a-values larger sample sizes are required when using ANCILLES than when using LOGIST, especially for the lower levels of overlap. For b-values, however, the reverse is true. Larger sample sizes are required for linking LOGIST b-values than for linking ANCILLES b-values.

Although c-values are not linked, they are averaged, and it is important to consider the quality of c-values used for an item pool. Even if good estimates of the a- and b-values are obtained from the linking procedure, poor c-values may lower the quality of the item pool. It is clear from this study that the ANCILLES program yields considerably better c-value correlations than the LOGIST program. This result, however, is probably an artifact of the restrictions placed on the c-values by LOGIST.

One last comment can be made regarding these two estimation procedures. The LOGIST estimates tended to be mildly subject to drift during linking of a-values, while the ANCILLES estimates were not.

Comparison of the Results for the 1PL and 3PL Models

The linking of b-values using the 1PL model was clearly superior to the linking of b-values using the 3PL model, except for the 2000 sample case. For the five item overlap level a sample of 2000 was needed before the correlations for the 3PL b-values were as high as the correlation for the 1PL b-values for the 100 sample case. The difference in sample size requirements became smaller as overlap increased, but it is clear that linking 3PL b-values requires greater sample sizes than the linking of 1PL b-values.

Because the 1PL model does not have discrimination and guessing parameters, the linking of a-values and c-values using the two models cannot be compared. If discrimination and guessing parameters are needed or desired, of course, it is clear that the linking of the estimates of these parameters must be done using the 3PL model. If only b-values are needed, then the linking of the 1PL b-values yields superior results.

Summary and Conclusions

The purpose of this study was to investigate the properties of available procedures for linking item parameter estimates. The properties investigated included sample size requirements, overlap requirements, and drift. From the analyses of these properties reported, the following conclusions were reached.

When large sample sizes were employed, the use of LOGIST a-values with any of the procedures appeared to yield adequate results. The same was true for the linking of b-values using the 3PL model. The best combination for linking b-values was the use of LOGIST b-values with any of the linking procedures. When small sample sizes were used for the linking of a-values the best results were obtained using the major axis and least square procedures with either the ANCILLES or LOGIST a-values. Again, the results were the same for small sample size linking of b-values. The linking results using the small sample sizes were not as satisfactory as the results obtained using large sample sizes. The small sample sizes appeared to be inadequate for both a-value and b-value linking using the 3PL model. For the linking of 1PL b-values a sample size of 100 appeared adequate.

Level of overlap did not seriously affect the results of linking for any of the procedures. However, it did appear that five items probably was not adequate. For the 15 and 25 item overlaps the results were quite similar, indicating that 15 items is probably sufficient overlap. With either 15 or 25 item overlap levels the results were as reported above. For the five item overlap the results were the same except the major axis method tended to yield less satisfactory results.

Based on the above conclusions the following recommendations were made. For best results an overlap of 15 items is probably best. An overlap of 25 items yields adequate results, but is probably impractical in many applications. At the 15 item overlap level a sample size of 2000 is probably needed for stable linking of a- and b-values when the

3PL model is used, although for LOGIST a-values 1000 is perhaps sufficient. For the 3PL model the LOGIST program appears to yield the best overall results. With a sample size of 2000 any of the procedures will probably yield adequate results. For the 1PL procedure a sample size of 100 to 300 will probably yield adequate results, assuming accurate parameter estimates.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Croll, P. R. and Urry, V. W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models. Washington, D.C.: Research Section, Personnel Research and Development Center, U.S. Civil Service Commission, June, 1978.

Ireland, C. M. An application of the Rasch one-parameter logistic model to individual intelligence testing in a tailored testing environment. Unpublished doctoral dissertation. University of Missouri-Columbia, 1976.

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14(2), 139-160.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M. D. Ability estimation and item calibration using the one- and three-parameter logistic models: A comparative study. (Research Report 77-1). Columbia, MO.: University of Missouri, November 1977 (NTIS No. AD A047943).

Rentz, R. R. Monitoring the quality of an item-pool calibrated by the Rasch model. Paper presented at the Annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.

Rentz, R. R. and Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14(2), 161-180.

Snedecor, G. W. and Cochran, W. G. Statistical methods. Ames, Iowa: The Iowa State University Press, 1967.

Wood, R. L., Wingersky, M. S. and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum RM-76-6). Princeton, N.J.: Educational Testing Service, June 1976.

Woodcock, R. W. Woodcock reading mastery tests. Circle Pines, Minn.: American Guidance Service, 1972.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14(2), 97-116.

Wright, B. D. and Panchapakesan, N. A procedure for sample free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Navy

1 Dr. Jack R Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Chief of Naval Education and Training
Liason Office
Air Force Human Resource Laboratory
Flying Training Division
WILLIAMS AFB, AZ 85224

1 CDR Mike Curran
Office of Naval Research
800 N. Quincy St.
Code 270
Arlington, VA 22217

1 Dr. Richard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Henry M. Halff
Department of Psychology,C-009
University of California at San Diego
La Jolla, CA 92093

Navy

1 Dr. Patrick R. Harrison
Psychology Course Director
LEADERSHIP & LAW DEPT. (7b)
DIV. OF PROFESSIONAL DEVELOPMMENT
U.S. NAVAL ACADEMY
ANNAPOLIS, MD 21402

1 CDR Charles W. Hutchins
Naval Air Systems Command Hq
AIR-340F
Navy Department
Washington, DC 20361

1 CDR Robert S. Kennedy
Head, Human Performance Sciences
Naval Aerospace Medical Research Lab
Box 29407
New Orleans, LA 70189

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr. James McBride
Navy Personnel R&D Center
San Diego, CA 92152

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

Navy

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington, DC 20390

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 LT Frank C. Petho, MSC, USN (Ph.D)
Selection and Training Research Division
Human Performance Sciences Dept.
Naval Aerospace Medical Research Laborat
Pensacola, FL 32508

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

Navy

1 Dr. Worth Scanland, Director
Research, Development, Test & Evaluation
N-5
Naval Education and Training Command
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Ronald Weitzman
Code 54 WZ
Department of Administrative Sciences
U. S. Naval Postgraduate School
Monterey, CA 93940

1 Dr. Robert Wisher
Code 309
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Dexter Fletcher
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria,VA 22333

1 Dr. Michael Kaplan
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 DR. JAMES L. RANEY
U.S. ARMY RESEARCH INSTITUTE
5001 EISENHOWER AVENUE
ALEXANDRIA, VA 22333

1 Mr. Robert Ross
U.S. Army Research Institute for the Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Army

1 Dr. Robert Sasmor
U. S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Commandant
US Army Institute of Administration
Attn: Dr. Sherrill
FT Benjamin Harrison, IN 46256

1 Dr. Frederick Steinheiser
Dept. of Navy
Chief of Naval Operations
OP-113
Washington, DC 20350

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Air Force

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Research and Measurment Division
Research Branch, AFMPC/MPCYPR
Randolph AFB, TX 78148

1 Dr. Malcolm Ree
AFHRL/MP
Brooks AFB, TX 78235

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

Marines

1 H. William Greenup
Education Advisor (E031)
Education Center, MCDEC
Quantico, VA 22134

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Major Michael L. Patrow, USMC
Headquarters, Marine Corps
(Code MPI-20)
Washington, DC 20380

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

CoastGuard

1 Mr. Thomas A. Warm
U. S. Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Other DoD

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. William Graham
Testing Directorate
MEPCOM/MEPCT-P
Ft. Sheridan, IL 60037

1 Military Assistant for Training and
Personnel Technology
Office of the Under Secretary of Defense
for Research & Engineering
Room 3D129, The Pentagon
Washington, DC 20301

1 Dr. Wayne Sellman
Office of the Assistant Secretary
of Defense (MRA & L)
2B269 The Pentagon
Washington, DC 20301

1 DARPA
1400 Wilson Blvd.
Arlington, VA 22209

Civil Govt

1 Dr. Andrew R. Molnar
Science Education Dev.
and Research
National Science Foundation
Washington, DC 20550

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

1 1 psychological research unit
Dept. of Defense (Army Office)
Campbell Park Offices
Canberra ACT 2600, Australia

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Capt. J. Jean Belanger
Training Development Division
Canadian Forces Training System
CFTSHQ, CFB Trenton
Astra, Ontario KOK 1B0

1 CDR Robert J. Biersner
Program Manager
Human Performance
Navy Medical R&D Command
Bethesda, MD 20014

1 Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Tel Aviv, Ramat Aviv 69978
Israel

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Liaison Scientists
Office of Naval Research,
Branch Office , London
Box 39 FPO New York 09510

1 Col Ray Bowles
800 N. Quincy St.
Room 804
Arlington, VA 22217

Non Govt

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 DR. C. VICTOR BUNDERSON
WICAT INC.
UNIVERSITY PLAZA, SUITE 10
1160 SO. STATE ST.
OREM, UT 84057

1 Dr. John B. Carroll
Psychometric Lab
Univ. of No. Carolina
Davie Hall 013A
Chapel Hill, NC 27514

1 Charles Myers Library
Livingstone House
Livingstone Road
Stratford
London E15 2LJ
ENGLAND

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester, NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. William E. Coffman
Director, Iowa Testing Programs
334 Lindquist Center
University of Iowa
Iowa City, IA 52242

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

Non Govt

1 Dr.,Fritz Drasgow
Yale School of Organization and Manageme
Yale University
Box 1A
New Haven, CT 06520

1 Dr. Mavin D. Dunnette
Personnel Decisions Research Institute
2415 Foshay Tower
821 Marguette Avenue
Mineapolis, MN 55402

1 Mike Durmeyer
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Benjamin A. Fairbank, Jr.
McFann-Gray & Associates, Inc.
5825 Callaghan
Suite 225
San Antonio, Texas 78228

1 Dr. Leonard Feldt
Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville, MD 20850

1 Univ. Prof. Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Non Govt

1 Professor Donald Fitzgerald
University of New England
Armidale, New South Wales 2351
AUSTRALIA

1 Dr. Edwin A. Fleishman
Advanced Research Resources Organ.
Suite 900
4330 East West Highway
Washington, DC 20014

1 Dr. John R. Frederiksen
Bolt Beranek & Newman
50 Moulton Street
Cambridge, MA 02138

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

1 Dr. Ron Hambleton
School of Education
University of Massechusetts
Amherst, MA 01002

1 Dr. Chester Harris
School of Education
University of California
Santa Barbara, CA 93106

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Library
HumRRO/Western Division
27857 Berwick Drive
Carmel, CA 93921

Non Govt

1 Dr. Steven Hunka
Department of Education
University of Alberta
Edmonton, Alberta
CANADA

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle, WA 98105

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Professor John A. Keats
University of Newcastle
AUSTRALIA 2308

1 Mr. Marlin Kroger
1117 Via Goleta
Palos Verdes Estates, CA 90274

1 Dr. Michael Levine
Department of Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat 23
9712GC Groningen
Netherlands

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

1 Dr. Gary Marco
Educational Testing Service
Princeton, NJ 08450

Non Govt

1 Dr. Scott Maxwell
Department of Psychology
University of Houston
Houston, TX 77004

1 Dr. Samuel T. Mayo
Loyola University of Chicago
820 North Michigan Avenue
Chicago, IL 60611

1 Professor Jason Millman
Department of Education
Stone Hall
Cornell University
Ithaca, NY 14853

1 Bill Nordbrock
Instructional Program Development
Building 90
NET-PDCD
Great Lakes NTC, IL 60088

1 Dr. Melvin R. Novick
356 Lindquist Center for Measurment
University of Iowa
Iowa City, IA 52242

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 Dr. James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

1 MR. LUIGI PETRULLO
2431 N. EDGEWOOD STREET
ARLINGTON, VA 22207

1 DR. DIANE M. RAMSEY-KLEE
R-K RESEARCH & SYSTEM DESIGN
3947 RIDGEMONT DRIVE
MALIBU, CA 90265

Non Govt

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Mark D. Reckase
Educational Psychology Dept.
University of Missouri-Columbia
4 Hill Hall
Columbia, MO 65211

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washington, DC 20007

1 Dr. Leonard L. Rosenbaum, Chairman
Department of Psychology
Montgomery College
Rockville, MD 20850

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

1 PROF. FUMIKO SAMEJIMA
DEPT. OF PSYCHOLOGY
UNIVERSITY OF TENNESSEE
KNOXVILLE, TN 37916

1 DR. ROBERT J. SEIDEL
INSTRUCTIONAL TECHNOLOGY GROUP
HUMRRO
300 N. WASHINGTON ST.
ALEXANDRIA, VA 22314

Non Govt

1 Dr. Kazuo Shigemasu
University of Tohoku
Department of Educational Psychology
Kawauchi, Sendai 980
JAPAN

1 Dr. Edwin Shirkey
Department of Psychology
University of Central Florida
Orlando, FL 32816

1 Dr. Robert Smith
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903

1 Dr. Richard Snow
School of Education
Stanford University
Stanford, CA 94305

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven, CT 06520

1 DR. PATRICK SUPPES
INSTITUTE FOR MATHEMATICAL STUDIES IN
THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CA 94305

1 Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

1 Dr. Brad Sympson
Psychometric Research Group
Educational Testing Service
Princeton, NJ 08541

Non Govt

1 Dr. Kikumi Tatsuoka
Computer Based Education Research
Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

1 Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

1 Dr. Robert Tsutakawa
Department of Statistics
University of Missouri
Columbia, MO 65201

1 Dr. J. Uhlaner
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

1 Dr. Howard Wainer
Division of Psychological Studies
Educational Testing Service
Princeton, NJ 08540

1 Dr. Phyllis Weaver
Graduate School of Education
Harvard University
200 Larsen Hall, Appian Way
Cambridge, MA 02138

1 Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044

1 Wolfgang Wildgrube
Streitkraefteamt
Box 20 50 03
D-5300 Bonn 2
WEST GERMANY

END

DATE
FILMED
12-81

DTIC